# Mining call detail records to reconstruct global urban mobility patterns for large scale emissions calculation

Manon Seppecher

▶ **To cite this version:**

Manon Seppecher. Mining call detail records to reconstruct global urban mobility patterns for large scale emissions calculation. Infrastructures de transport. Université de Lyon, 2022. English. NNT : 2022LYSET002 . tel-03703230

**HAL Id: tel-03703230**
**https://theses.hal.science/tel-03703230**

Submitted on 23 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

N°d'ordre NNT : 2022LYSET002

**THÈSE de DOCTORAT DE L'UNIVERSITÉ DE LYON**
opérée au sein de
**l'Ecole Nationale des Travaux Publics de l'Etat**

**École Doctorale** N° 162
**MEGA (Mécanique, Energétique, Génie Civil et Acoustique)**

**Spécialité / discipline de doctorat** : Génie Civil

Soutenue publiquement le 24/01/2022, par :
**Manon Seppecher**

---

# Exploration de données de téléphonie mobile pour la reconstruction de patterns globaux de mobilité urbaine pour le calcul d'émission à large échelle

---

Devant le jury composé de :

| | | | |
|---|---|---|---|
| Hassas, Salima | Prof. | Univ. Lyon | Présidente |
| Fiore, Marco | Prof. | IMDEA Network Institute | Rapporteur |
| Geroliminis, Nikolas | Prof. | EPFL | Rapporteur |
| Canudas-De-Wit, Carlos | Prof. | Grenoble INP | Examinateur |
| André, Jean-Marc | PhD. | Citepa | Examinateur |
| Leclercq, Ludovic | Prof. | Univ. Eiffel, ENTPE | Directeur de thèse |
| Lejri, Delphine | ITPE | ENTPE, Univ. Eiffel | Co-directrice de thèse |
| Furno, Angelo | CR | ENTPE, Univ. Eiffel | Invité |

Thèse co-encadrée par le CITEPA, référente: Thamara Vieira Da Rocha

N°d'ordre NNT : 2022LYSET002

**A THESIS OF THE UNIVERSITY OF LYON**
Prepared at
**Ecole Nationale des Travaux Publics de l'Etat**

**Doctoral School** N° 162
**MEGA (Mechanics, Energy, Civil Engineering and Accoustics)**

To obtain the graduation of
**PhD in Civil Engineering**

Defended in public on 01/24/2022, by:
**Manon Seppecher**

---

# Mining call detail records to reconstruct global urban mobility patterns for large scale emissions calculation

---

In front of the following examination committee:

| | | | |
|---|---|---|---|
| Hassas, Salima | Prof. | Univ. Lyon | Committee chair |
| Fiore, Marco | Prof. | IMDEA Network Institute | Reviewer |
| Geroliminis, Nikolas | Prof. | EPFL | Reviewer |
| Canudas-De-Wit, Carlos | Prof. | Grenoble INP | Examiner |
| André, Jean-Marc | PhD. | Citepa | Examiner |
| Leclercq, Ludovic | Prof. | Univ. Eiffel, ENTPE | Supervisor |
| Lejri, Delphine | ITPE | ENTPE, Univ. Eiffel | Supervisor |
| Furno, Angelo | CR | ENTPE, Univ. Eiffel | Guest |

Thèse co-encadrée par le CITEPA, référente: Thamara Vieira Da Rocha

# Preface

At the end of this three-year-long experience, I would like to thank those who contributed directly or indirectly to the completion of this thesis.

My warm appreciation and gratitude first go to my Ph.D. director, Prof. Ludovic Leclercq, for being such a receptive and supportive director. Our collaboration has been extremely instructive and enriching, and I expect a lot from working together on new research subjects. My sincere thanks also go to my co-director Delphine Lejri. Our exchanges have always helped me take a step back from my work. She has also often helped me keep my confidence, and I am grateful to her for that. I am very thankful for the support of my supervisor Angelo Furno, who always made sure to ask me challenging questions about my work. Throughout these three years, his rigor and requirements have certainly made this thesis better.

Besides, I would like to warmly thank all my hierarchy at Citepa, without whom this thesis would never have started. First, I would like to express my deepest gratitude to my supervisor Thamara Vieira Da Rocha. Working with her was highly satisfying. I believe we have taken the GCBD project a step forward, and I am proud of that. I sincerely thank my team leader, Irving Tapia Villarreal, for being constantly encouraging, from Paris to Medellin, and my department head, Jean-Marc André, for his trust and support to my work. Working on their team has been a source of important satisfaction. I also would like to address my appreciation to Jérôme Boutang, director of Citepa, for coming up with such an ambitious research project and entrusting me with this subject. Finally, I would like to express my gratitude to Bertrand Bessagnet, my former department head at Citepa. I appreciated his interest in my work very much.

My thanks also go to my colleagues, both at Citepa and at the Licit laboratory. I have learned much from them as a scientist, an engineer, and a person, and the memories made in Paris and Lyon are dear to me. Most notably, I am very grateful to my ex-colleague at Licit lab and friend Sergio Batista for sharing with me his research questions and helping me solve mine. I also thank Felipe Troncoso-Lamaison, my colleague at Citepa, for his strong curiosity about my work and always accurate questions. I am confident that the GCBD project is safe with him and Thamara. How not to mention Anne-Christine and Sonia, the lab assistants, for their precious help throughout these three years, Anne-Christine for her constant and unfailing help in organizing my travels to Paris, Sonia for her patient and positive support in all my administrative procedures at ENTPE. Finally, I would like to thank my former office mates Sergio, Guilhem and Humberto, for the small and big talks; Elise, for the nights out from Lyon to Washington; Ruiwei and Anna, for all the sportive and artistic activities we shared. I am grateful for those friendships earned along the way.

I am fortunate to be surrounded by wonderful friends. Even from afar, they have contributed to my happiness during these three years. I would especially like to thank my dear friends Lauriane and Simon for often opening their door to me during my Parisian stays, I felt at home with them.

Most importantly, from the bottom of my heart, I would like to thank my lover Jordan, my little sister Lisa, my parents, and my grandparents. I look up to who they are, and I am immensely grateful for their love, encouragement, and unfailing support.

# Abstract

Urban air quality is a crucial environmental and health issue. This issue is especially central in the fight against climate change and adapting cities to this change. Road traffic and chronic congestion contribute significantly to atmospheric emissions in urban areas. Therefore, joint monitoring of road traffic and related emissions is an essential support for urban public decision-making. And beyond this kind of procedure, public authorities need methods for evaluating transport policies according to environmental criteria.

Coupling dynamic traffic models with traffic-related emission estimation models is a suitable response to this need. However, integrating this solution into decision support tools requires characterizing urban mobility in near to real-time. In some cities, major roads are equipped with loop detectors. They provide the data needed for these analyses. However, the installation of such sensors is costly and far from being widespread. In a context of continuously increasing cell phone usage, the data they generate, and in particular Call Detail Records, appear to be an interesting alternative to traditional data. These data are rich, massive, with high penetration rates, and available worldwide. They have already supported critical mobility analysis, such as origin-destination matrices estimations. However, their use for systematic traffic characterization has remained limited. It is mainly due to the low spatial resolution of these data and their temporal sampling rate, which is sensitive to communication behaviors.

This Ph.D. thesis contributes to developing a methodological framework for estimating the traffic variables needed to calculate air emissions: total travel distances and average traffic speeds. One of the main issues is estimating total travel distances, despite the different data biases. We face two main problems: 1. the integration of different urban mobility profiles in the analysis; 2. the reconstruction of complete individual and collective mobility patterns, despite the fragmented nature of the data. In response to these two issues, an essential contribution of this thesis is to articulate methods for classifying individuals with two different approaches for reconstructing users' mobility according to their mobility profiles. Another fundamental issue is to estimate average traffic speeds when the quality of the data does not allow to identify the routes taken, nor to estimate reliable individual travel time and speeds. A key contribution of this thesis is the development of an innovative method for estimating traffic speeds, designed to handle the temporal biases of cell phone data, and more generally, of any positioning data depending on the user's activity. This approach is based on the fusion of a large amount of travel data and an aggregated representation of the road network consistent with the scales used for mobility reconstruction. The proposed methodological framework relates these different methods coherently and presents a complete modeling and data processing process.

This thesis is part of a larger Research and Development project led by Citepa, a French state operator for the Ministry of the Environment. The project, entitled Green City Big Data, focuses on developing a decision support tool based on cell phone data and allowing the evaluation of public transport policies. This project relies on a partnership with the Colombian communication provider CLARO, which has shared with Citepa large amounts of anonymized CDR data on which most of this research is based.

# Resumé

La qualité de l'air en milieu urbain est un enjeu sanitaire crucial. Par ses effets induits sur le climat, cet enjeu est aussi au cœur de la lutte contre le changement climatique et de la question de l'adaptation des villes à ce changement. Le trafic routier et les phénomènes de congestion chroniques contribuent de manière significative aux émissions atmosphériques dans les zones urbaines. Par conséquent, la surveillance conjointe du trafic routier et des émissions qu'il génère constitue un support essentiel de la décision publique en matière d'adaptation. Au-delà de simples procédures de suivi, les pouvoirs publics ont besoin de méthodes d'évaluation des politiques de transport selon des critères environnementaux.

Le couplage de modèles dynamiques de trafic avec des modèles d'émissions constitue une réponse adaptée à ce besoin. Cependant, l'intégration de tels modèles à des outils d'aide à la décision nécessite de caractériser la mobilité urbaine en temps quasi réel. Pour les axes majeurs de certaines villes, les boucles de comptage de véhicules fournissent les données nécessaires à ces analyses. Mais l'installation de tels capteurs est coûteuse et loin d'être généralisée. Or dans un contexte d'augmentation continue des usages liés à la téléphonie mobile, les données produites par ces usages, et en particulier les statistiques d'appel (données CDR), apparaissent comme une alternative intéressante aux données traditionnelles. Ces données sont riches, massives, avec des taux de pénétration élevés, et disponibles partout dans le monde. Elles ont déjà fait l'objet d'études importantes pour l'analyse de la mobilité, comme l'estimation de matrices Origine-Destination. Néanmoins, leur utilisation pour la caractérisation systématique du trafic est restée relativement limitée. Cela s'explique notamment par la faible résolution spatiale de ces données, et leur taux d'échantillonnage temporel qui est sensible aux comportements de communication.

Cette thèse de doctorat contribue au développement d'un cadre méthodologique permettant l'estimation des variables de trafic nécessaires au calcul d'émissions atmosphériques: les distances totales parcourues et les vitesses moyennes de trafic. L'un des principaux enjeux est d'estimer des distances totales parcourues, malgré les différents biais des données. Pour cela, nous nous heurtons à deux problématiques essentielles : 1. l'intégration dans l'analyse des différents profils de mobilité urbaine ; 2. la reconstruction des schémas de mobilité individuels et collectifs complets, en dépit du caractère parcellaire des données. En réponse à ces deux problématiques, une contribution importante de cette thèse est d'articuler des méthodes de classification des individus avec deux approches distinctes de reconstruction de la mobilité des utilisateurs, adaptées aux différents profils de mobilité. L'estimation de vitesses moyennes de trafic soulève un second enjeu fondamental, car la qualité des données ne permet ni d'identifier les itinéraires empruntés, ni d'estimer des vitesses individuelles fiables. Une apport clé de cette thèse est le développement d'une méthode innovante d'estimation des vitesses de trafic, conçue pour gérer les biais temporels des données de téléphonie mobile, et plus généralement, de toute donnée de positionnement dépendant de l'activité de l'utilisateur. Cette approche est basée sur la fusion d'une large quantité de données de déplacements, et sur une représentation agrégée du réseau routier cohérente avec les échelles de reconstruction de la mobilité. Le cadre méthodologique proposé articule ces différentes méthodes de façon cohérente et propose un processus complet de modélisation et de traitement des données.

Cette thèse s'inscrit dans le cadre d'un projet de Recherche et Développement plus vaste, porté par le Citepa, opérateur d'Etat français pour le Ministère de la Transition Ecologique. Le projet, intitulé Green City Big Data, porte sur le développement d'un outil d'aide à la décision basé sur des données de téléphonie mobile et permettant l'évaluation de politiques publiques de transports. Ce projet s'appuie sur un partenariat avec l'operateur de communication colombien CLARO, qui a partagé avec le Citepa de grandes quantités de données CDR anonymes sur lesquelles l'essentiel de ces travaux de recherche sont basés.

# Resumé long

## Introduction

En ce début de XXIe siècle, la qualité de l'air et l'adaptation au dérèglement climatique représentent deux enjeux majeurs pour les villes. Celles-ci visent donc à se doter d'outils de contrôle efficaces pour suivre l'évolution de la qualité de l'air, et pour évaluer et adapter leur politiques de transports de façon à réduire leur impact environnemental et à lutter de façon efficace contre les phénomènes de pollution. L'adoption de tels outils et procédures est notamment cruciale pour les villes des pays en développement, car elles sont confrontées à une croissance démographique et urbaine plus forte que les pays occidentaux, et à des situations chroniques de congestion du trafic et de pollution de l'air. Dans ce contexte, et alors que les capteurs traditionnels de surveillance du trafic et de la qualité de l'air manquent, il est nécessaire de concevoir de nouvelles méthodes et de développer des outils moins coûteux pour évaluer les politiques publiques en matière de transport et de qualité urbaine.

Le développement de la téléphonie mobile et l'accessibilité des statistiques d'appels (données CDR ou Call Detail Records) ont ouvert le champ à une utilisation de ces données pour l'analyse de la mobilité. Pourtant, peu d'études ont pour l'instant évalué l'intérêt de ces données pour la modélisation du trafic routier et des émissions associées. C'est pourquoi le Citepa, association a but non lucratif et opérateur de l'Etat français pour le Ministère de la Transition Ecologique, a initié le projet de recherche et de développement Green City Big Data. Ce projet a pour objectif de démontrer l'utilité des données CDR pour l'évaluation environnementale des politiques de transport en milieu urbain. Il s'appuie sur un partenariat étroit avec l'opérateur de téléphonie mobile américain CLARO, qui fournit d'importants jeux de données anonymisées sur lesquelles sont conduits les travaux de recherche. La thèse présentée ici a été financée par le Citepa dans le cadre du projet Green City Big Data, selon une convention Cifre, et a bénéficié de l'accès à ces données. Elle porte sur l'estimation des variables nécessaires au calcul des émissions atmosphériques du trafic routier en milieu urbain. L'objectif est de proposer des méthodes de traitements de données pour estimer de deux variables principales: le volume et la vitesse moyenne du trafic. La question du partage modal, également indispensable à l'estimation des émissions atmosphériques relatives au trafric, était hors du périmètre de recherche de cette thèse, mais fait l'objet aussi l'objet d'étude dans le cadre du projet GCBD.

Les données CDR sont issues de la collecte des événements de télécommunication générés par les utilisateurs du service: appels, SMS, connexions internet. Ces données sont sensibles aux habitudes de communication des utilisateurs, et comportent donc des biais temporels. Leur résolution spatiale, calquée sur le réseau d'antennes, est également très limitée. Pour ces raisons, et bien qu'elles aient déjà fait l'objet d'études approfondies pour l'analyse de la mobilité, ces données sont *a priori* peu adaptées à la caractérisation du trafic routier.

En la matière, la littérature présente des limites significatives. L'estimation de variables telles que la vitesse, les temps de déplacement, les distances parcourues ou les flux

régionaux a été négligée. La recherche s'est surtout concentrée sur l'estimation de matrices origine-destination [Iqbal et al., 2014, Alexander et al., 2015, Toole et al., 2015] et sur l'identification de chaînes d'activité individuelles. Cette littérature se base souvent sur des sous-échantillons de population favorables à une analyse des données (utilisateurs très actifs, réguliers, résidents), omettant de caractériser d'autres individus et donc d'autres formes de mobilité. Ces travaux ignorent aussi souvent l'impact des biais temporels des données sur les résultats, qui pourtant induisent une représentation partielle de la mobilité [Zhao et al., 2021, Hoteit et al., 2017, Chen et al., 2018].

Nous nous attachons donc à développer des méthodes d'estimation des variables du trafic basées sur une caractérisation plus globale de la mobilité. Cela passe par la prise en compte d'un éventail de profils de mobilité différents, et par la reconstruction des informations partielles de déplacement.

Cette thèse est divisée en trois parties. La première partie présente les données, le territoire étudié, et le traitement préliminaire des données qui soutient la mise en œuvre de la chaîne de modélisation proposée dans cette thèse. La deuxième partie de cette thèse se concentre sur l'analyse de la mobilité urbaine, avec pour objectif la reconstruction du volume de trafic pour deux catégories différentes de la population, les utilisateurs réguliers et les utilisateurs non réguliers. La troisième et dernière partie de la thèse rassemble des contributions importantes pour le calcul des émissions atmosphériques. Le Chapitre 5 expose une nouvelle méthode d'estimation de la vitesse du trafic (nécessaire au calcul des facteurs d'émission) à partir d'un type générique de données sensibles aux rythmes de communication des utilisateurs. Le Chapitre 6 lui présente l'articulation des différentes méthodes développées dans cette thèse au sein d'une chaîne globale de modélisation et de traitement de données. Un dernier court chapitre conclut se manuscrit en en résumant les contributions et en énumérant les principales perspectives de recherche prolongeant ce travail.

# Partie I

Cette première partie du manuscrit est consacrée à la présentation de l'étude de cas et des données, et à la conduite d'études préliminaires sur la population et du réseau routier. Ces analyses soutiendrons le développement de l'ensemble de la chaîne de modélisation proposée dans cette thèse.

### Chapitre 1: Présentation de l'étude de cas

Dans ce chapitre, nous présentons le périmètre étudié, les données accessibles, et l'impact de leur résolution spatiale sur notre analyse.

Notre étude de cas porte sur la ville de Santiago de Cali, en Colombie. Santiago de Cali est la troisième ville la plus peuplée de Colombie. Elle est située dans le département de Valle del Cauca, dans le sud-ouest du pays. L'agglomération est divisée en 22 districts urbains et 15 districts ruraux. La municipalité de Cali est fortement liée à deux municipalités voisines, Yumbo au nord et Jamundi au sud, qui sont incluses dans notre analyse. La zone urbaine de Cali s'étend sur 123 kilomètres carrés, tandis que la superficie de l'ensemble de la municipalité est de 569 kilomètres carrés. Selon le dernier recensement national [DANE], 2 172 527 personnes vivent dans la zone urbaine de Cali, tandis que 55 115 personnes vivent dans les districts ruraux. Un schéma illustre le découpage administratif, ainsi que la distribution de la population sur le territoire.

Dans la deuxième partie du chapitre, après avoir rappelé les propriétés particulières des données CDR, nous présentons le jeu de données fourni par CLARO en tant que partenaire du projet Green City Big Data. Avant le transfert, les données ont été compressées

par l'opérateur afin d'assurer la sécurité des données des utilisateurs. Le processus de compression consiste à supprimer tout événement de communication qui n'apporterait pas d'information sur la mobilité. Outre la limitation des informations individuelles partagées, ce processus de compression présente l'avantage de réduire significativement la taille des données. Le jeu de données reçu couvre une période de 23 mois, allant de janvier 2020 à novembre 2021. Seuls les trois premiers mois de données ont été exploités dans cette thèse. Pendant cette période, un total de 2 707 012 identifiants d'utilisateurs ont été observés sur la zone couverte, qui inclut la municipalité de Santiago de Cali, ainsi que les municipalités voisines de Yumbo et Jamundi. Un total de 440 stations de base couvre Cali, Yumbo et Jamundi, avec des densités inégales décrites dans le chapitre.

Le partitionnement de Voronoi de la zone selon le réseau d'antennes a pour résultat un ensemble de régions aux superficies disparates. Nous proposons de définir une échelle supérieure d'analyse faite de régions de superficies et de tailles de population homogènes. Dans la dernière partie de ce chapitre, nous développons une méthode simple pour générer un nouveau partitionnement spatial en agrégeant les stations de base. Dans un premier temps, l'algorithme agrège les cellules selon le nombre d'utilisateurs identifiés comme résidents de cette cellule. Le but de cette phase est de s'assurer que toutes les régions ont un sous-échantillon de population suffisamment grand pour garantir une représentativité statistique fiable des informations tirées de cet échantillon. Dans un deuxième temps, l'algorithme agrège les régions sur la base de leur superficie. Cette phase permet d'assurer que les régions ont des superficies similaires. Le seuil de population est fixé à 400 individus tandis que différents seuils de superficie sont explorés. Ce double objectif permet d'agréger des cellules en zone rurale, où les stations de base sont peu nombreuses et la population locale faible, et en zone urbaine, où les stations de base couvrent de petites surfaces avec une forte densité de population. Le zonage généré avec un critère de surface de 100 ha définit la résolution spatiale utilisée pour l'estimation des variables dans le reste de la thèse.

### Chapitre 2: Analyses préliminaires de l'échantillon d'utilisateurs et du réseau routier

Ce deuxième chapitre rassemble deux études préliminaires essentielles à l'application de notre chaîne de traitement de données et de modélisation. La première porte sur une analyse de la population, la deuxième sur une analyse du réseau routier. Toutes deux fournissent les bases d'une analyse complète et exhaustive de la mobilité.

Dans la première partie de ce chapitre, nous nous interrogeons sur les moyens de caractériser la mobilité de l'ensemble de la population captée par les données CDR, en dépit d'habitudes de communication et de fréquentation de la zone d'étude variées. De nombreuses études bibliographiques basées sur les données CDR se sont concentrées, par souci de simplicité, sur la mobilité des résidents des zones étudiés. Dans ces circonstances, les conclusions tirées (matrices OD) ne peuvent être représentatives que de cette part de la population urbaine, mais ne rendent pas compte de la mobilité globale, qui comprend d'autres types d'utilisateurs. Un enjeu critique de notre méthodologie est donc d'identifier, à partir des données CDR, différents profils de mobilité et d'associer à chacun un facteur de redressement et d'expansion juste.

Nous proposons d'abord de catégoriser les utilisateurs détectés selon différents profils de mobilité, avant de développer pour chacun une méthode d'estimation des facteurs d'expansion associés. Trois profils de mobilité sont sélectionnés: les résidents, les navetteurs et les visiteurs. Pour catégoriser les utilisateurs selon l'un de ces profils, nous utilisons une approche de classification simple, où les limites des clusters sont déterminées par un ensemble de règles paramétriques. Le nombre de jours observé dans la zone, le nombre de jours ouvré observé dans la zone, le nombre de nuits observé dans la zone, le séjour le plus court

(en nombre de jours consécutifs) observé sur la période historique sont choisies comme variables discriminantes. Les seuils de classification sont calibrés de façon à respecter certaines constantes macroscopiques issues des recensements et enquêtes de mobilités.

Lorsque le lieu de résidence des individus de l'échantillon peut être estimé, on rapporte en général la taille de l'échantillon local aux données de recensement pour calculer des facteurs d'expansion et de redressement. En revanche, la littérature existante dans le domaine de la classification des utilisateurs ne se penche pas sur la question du redressement des échantillons d'utilisateurs dont le lieu de résidence ne peut être déterminé (utilisateurs non résidents). Nous proposons donc une stratégie pour chacun de ces profils d'utilisateurs (navetteurs, visiteurs) sur la base de données de recensement, d'enquête de mobilité et de tourisme. Une limite de cette approche est que les données d'enquêtes ne sont pas partagées selon des formats standardisés, contrairement aux données de recensement. Cela complexifie la méthode et limite sa reproductibilité sur d'autres territoires.

Par manque de données complémentaires, il est difficile de valider la classification des utilisateurs. L'exploration des propriétés de présence et de mobilité des différents groupes d'utilisateurs peut toutefois confirmer que la classification est cohérente à un niveau agrégé. Nous extrayons les taux de présence moyens par classe d'utilisateurs sur une période historique de deux mois et observons qu'ils sont conformes à nos attentes. Des rythmes hebdomadaires et saisonniers clairs sont en particulier identifiés. Pour finir, nous calculons les facteurs d'expansion et discutons les résultats. Plusieurs explications sont avancées pour justifier les facteurs d'expansion faibles calculés pour certaines régions ou certaines catégories d'utilisateurs.

La seconde partie de ce chapitre porte sur la caractérisation du réseau routier. Le but est de fournir des données préalables à la reconstruction de variables de trafic à partir de données CDR. Pour estimer les volumes de trafic, nous découplons flux régionaux et distances moyennes, et d'estimer ces deux variables séparément. L'estimation des flux régionaux nécessite d'identifier les trajectoires empruntées par les utilisateurs. La littérature sur ce sujet est limitée, et les méthodes de la littérature pour la reconstruction de trajectoires sont soit inadaptées aux données CDR, soit trop coûteuses. Nous proposons d'utiliser des travaux de la littérature [Batista et al., 2019] développés dans le contexte de la théorie MFD (Macroscopic Fundamental Diagram) pour conduire une analyse hors-ligne du réseau. Le travail de Batista et al. [2019], développé à partir d'un échantillonnage de trajets synthétiques, porte sur l'identification automatique des chemins dominants du réseau, et sur l'estimation des longueurs moyennes de parcours à l'échelle des régions. Cette analyse permet d'obtenir une connaissance préalable des itinéraires sur le réseau, pouvant servir de support à une assignation ultérieure des traces à un itinéraire (ou *map matching*). En permettant aussi d'estimer les longueurs moyennes de parcours, cette approche palie les limites de la résolution spatiale des données CDR.

Après avoir introduit la notion de chemin régional essentielle ici [Yildirimoglu et Geroliminis, 2014, Batista et al., 2019], nous présentons l'algorithme tel qu'implémenté dans la littérature. Il consiste à échantillonner un ensemble de couples origine-destination aléatoires, puis à appliquer à chaque couple un algorithme de calcule de plus court chemin (Dijkstra) pour identifier le chemin dominant correspondant. Nous proposons deux modifications de la méthode pour la rendre scalable à de grands réseaux.

La première modification s'appuie sur une propriété de l'algorithme de Dijkstra qui n'était pas exploitée dans la version originale de la méthode. Cette propriété est que l'algorithme calcule non seulement le plus court chemin d'un point $A$ à un point $B$, mais d'un point $A$ à tous les autres noeuds du graphe. Nous tirons parti de cette propriété pour limiter le nombre de couples origine-destination échantillonnés et le nombre d'exécutions de l'algorithme de Dijkstra tout en maintenant un haut niveau de caractérisation du réseau.

La deuxième adaptation de la méthode porte sur l'échantillonnage des nœuds d'origine

et de destination. Dans Batista et al. [2019], les auteurs recommandent d'échantillonner de nombreux couples origine-destination afin d'assurer une bonne couverture du réseau et donc une estimation précise des longueurs. Nous proposons de réduire drastiquement le nombre de nœuds candidats à l'échantillonnage, en nous concentrant sur les plus stratégiques, c'est à dire ceux localisés aux frontières des régions, par lesquels les flux entrant ou sortant doivent nécessairement transiter.

Nous discutons de l'impact de ces modifications sur les résultats et les temps de calculs, et application la nouvelle version de la méthode à notre étude de cas. Les résultats nous fournissent un aperçu précieux des caractéristiques spécifiques du réseau, et des données essentielles pour la suite du développement de notre chaîne de modélisation. Premièrement, nous avons identifié les chemins dominants entre une origine et une destination régionales. Ces chemins dominants serviront plus tard de base pour déduire les chemins empruntés par les utilisateurs, et donc pour estimer les flux de population appropriés. Deuxièmement, la connaissance des longueurs de parcours régionales, une fois croisée avec les flux, fournira une estimation des distances totales parcourues dans chaque région. La connaissance des longueurs de parcours régionales sera aussi précieuse pour l'estimation des vitesses de trafic régionales.

# Partie II

La deuxième partie de la thèse porte sur l'estimation des volumes de trafic à partir des données CDR.

La classification des utilisateurs menée dans le Chapitre 2 a permis de les catégoriser selon trois profils: résidents, navetteurs et visiteurs. Ces trois catégories peuvent être regroupées en deux catégories principales, les utilisateurs locaux (résidents et navetteurs) et les utilisateurs non locaux (visiteurs). Une différence essentielle entre ces individus est la profondeur de l'historique de données disponible. Alors que les visiteurs font des séjours courts, les résidents et les navetteurs sont observés régulièrement et ont des historiques de données beaucoup plus étendus. La disponibilité de ces historiques constitue une ressource importante pour interpréter et enrichir la mobilité partielle extraite des données CDR, puisqu'elle est largement contrôlée par un certain nombre d'habitudes qui la rendent très régulière, tant à l'échelle individuelle que collective. Néanmoins, certains individus, même locaux, ont une mobilité (personnelle ou professionnelle) plus erratique. Pour ceux-là, même la disponibilité d'un historique de mobilité fourni peut ne pas suffire à enrichir des observations qui ne se rapportent pas à la mobilité passée.

Dans cette partie, nous proposons donc deux méthodes distinctes, l'une conçue pour être appliquée aux utilisateurs ayant une mobilité régulière, l'autre à des utilisateurs ayant une mobilité irrégulière. Les utilisateurs identifiés comme étant des visiteurs sont d'emblée considérés comme irréguliers, du fait de leur faible profondeur historique. Pour les locaux (résidents et navetteurs), la distinction entre utilisateurs réguliers et utilisateurs irrégulier est introduite via la notion d'entropie [Song et al., 2010]. Sur la base de la distribution de l'entropie, les utilisateurs ayant une entropie aberrante sont considérés comme irréguliers.

### Chapitre 3: Assignation des déplacements à des itinéraires et estimation des flux par chemins régionaux

Ce chapitre se concentre sur des utilisateurs réguliers. On y propose une démarche méthodologique pour estimer à partir de données CDR les flux par chemins régionaux, ou *path flow* en anglais. La prise en compte du biais temporel des données est central: afin d'estimer des flux justes, il est nécessaire d'identifier des chaînes d'activité complètes, et de reconstruire les trajectoires des individus entre les activités consécutives.

Dans un premier temps, nous nous intéressons à l'enrichissement des chaînes d'activités quotidiennes, c'est à dire des séquences des lieux visités par un individu au cours d'une journée. Sur la base de l'historique de données de l'utilisateur, nous commençons par identifier les routines de mobilité sur la base de l'historique de données des utilisateurs. Pour un utilisateur donné, les plus riches de ses chaînes d'activité historiques sont clusterisées (avec l'algorithme DBSCAN) de façon à identifier différents motifs redondants. Ensuite, nous proposons une heuristique pour compléter les chaînes d'activité partielles sur la base des modèles fournis par les motifs historiques. Cette méthode aboutit à l'enrichissement des chaînes d'activités de 11%, et une augmentation satisfaisante des flux au cours de la journée. Cependant, nous constatons un enrichissement excessif de la mobilité la nuit et le week-end. Ces observations s'expliquent notamment par l'absence de distinction du jour de la semaine dans la construction des routines. D'autres biais de notre approche sont discutés et nous proposons des perspectives pour les prendre en compte dans de futures améliorations de la méthode.

La sensibilité aux événements de communication limite également l'identification des trajectoires individuelles entre les activités consécutives, et donc la reconstruction des flux de voyageurs. Plutôt que d'adopter les méthodes de *map matching* coûteuses de la littérature souvent développées pour des données plus riches que les données CDR, nous proposons une méthode de reconstruction des itinéraires efficace qui s'appuie sur la connaissance préalable des chemins dominants du réseau (cf. Chapitre 2). L'analyse des traces CDR les plus riches permet d'identifier des itinéraires alternatifs qui n'auraient pas été détectés par l'approche topologique. Ces itinéraires sont ajoutés à la base de chemins régionaux principaux. Cette base de données enrichie permet d'interpréter de façon efficace les traces de mobilité partielles: on peut associer à chaque déplacement observé l'itinéraire complet le plus vraisemblable. Dans le cas où aucune position intermédiaire ne caractérise un déplacement, nous appliquons une procédure d'assignation probabiliste, basée sur une l'analyse dynamique de la distribution des flux par chemins régionaux. La multiplication de ces flux par les distances régionales fournira les volumes de trafic régionaux.

Pour finir, ce chapitre comprend enfin une discussion importante sur les efforts de validation restant à conduire sur ces travaux, mais pour lesquels l'absence de données complémentaires constitue un obstacle sérieux.

### Chapitre 4: Reconstruction de la mobilité irrégulière

Si la plus part des voyageurs sont caractérisés par des motifs réguliers de déplacement urbain, certains ont des comportements plus erratiques. Par exemple, les visiteurs et certains locaux (chauffeurs de taxi ou livreurs) sont susceptibles d'avoir des traces de mobilité très variées d'un jour à l'autre. Dans la littérature, de telles traces de mobilité sont souvent ignorées car jugées aberrantes. Or, ces usagers peuvent être encore plus mobiles que les autres usagers réguliers et donc contribuer d'autant plus aux émissions atmosphériques: il nous faut donc les inclure dans notre analyse.

Ceci n'est pas trivial, car l'irrégularité des schémas de mobilité de ces utilisateurs limite l'utilisation des données historiques pour interpréter et enrichir les observations de mobilité. Cette irrégularité pose aussi la question de la représentativité de leur mobilité au sein de leur propre population: cette représentativité est discutable, du moins à l'échelle fine des chemins régionaux. Il semble donc à la fois difficile et peu légitime de chercher à reconstruire la mobilité des individus à cette échelle. En revanche, on peut supposer que les distances parcourues quotidiennement, elles, sont représentatives. Pour cette part de la population, c'est cette variable que nous allons chercher à estimer. Ce chapitre a donc pour objectif de proposer une méthode complémentaire d'estimation des distances parcourues, adaptée à une population non régulière. Il est divisé en deux sections principales, l'une exposant la méthodologie, l'autre présentant les résultats et une analyse d'application.

La méthode que nous proposons est basée sur une reconstruction collective de la mobilité à l'échelle de la ville de Cali. Plutôt que de chercher à enrichir la mobilité de façon individuelle comme dans le chapitre précédent, nous proposons une approche plus collective. A une échelle spatiale et temporelle macroscopique, la mobilité des utilisateurs aux rythmes de communication les plus faibles est supposée similaire à la mobilité des utilisateurs les plus actifs, dont les observations de mobilité peuvent être considérés comme complètes. Nous concentrons donc notre analyse des distances sur ce second sous-échantillon, avant d'étendre les conclusions à l'ensemble de l'échantillon, puis à la population globale qu'il représente.

Les niveaux de communication des utilisateurs sont mesurés selon un taux de couverture quotidien, qui quantifie la part d'une journée couverte par des événements de communication générés par l'utilisateur. Une analyse de sensibilité permet de définir le taux seuil au-delà duquel les observations de mobilité sont considérées comme complètes. Les utilisateurs dont le taux de couverture est supérieur au seuil fixé sont considérés comme représentatifs et leur distance quotidienne globale est calculée.

L'objectif de représentativité de la mobilité du sous-échantillon nous impose d'évaluer les distances parcourues à l'échelle de la ville. Les données de longueurs de parcours fournies par l'analyse topologique du réseau ne sont pas adaptées à cette échelle: un calcul complémentaire des distances parcourues est donc requis. Selon la taille du réseau d'antennes et sa granularité, le calcul systématique du plus court chemin entre les origines et les destinations observées peut être coûteux. Pour contourner cette étape, nous proposons une méthode hybride de calcul des distances, s'appuyant sur un calcul de plus court chemin pour les trajets les plus courts, et sur la notion de ratio de détour, développée dans la littérature, pour les longs trajets. Le ratio de détour mesure le rapport d'une distance avec détour à une distance de référence. Pour un couple origine-destination donné, nous définissons notre ratio de détour comme le rapport entre la distance du plus court chemin et la distance euclidienne. Une fois calibrée, la fonction de ratio de détour permet d'estimer de façon très efficace la distance de plus court chemin directement à partir de la distance euclidienne.

Dans une section consacrée aux résultats, nous commençons par illustrer la calibration de la fonction de ratio de détour, et discutons les résultats au regard du jeu de données utilisé. Ensuite, nous validons la méthode avec une approche expérimentale basée sur l'échantillonnage de trajectoires. Pour ces traces synthétiques, nous comparons les distances estimées via notre approche hybride avec la longueur réelle de la trajectoire. Nous montrons que les erreurs d'estimation sont limitées, et que les résultats produits sont beaucoup plus fins que lors de l'application d'une simple distance euclidienne, métrique adoptée dans certains travaux de la littérature. Ensuite, nous présentons les résultats de la mise en œuvre de cette méthodologie sur Santiago de Cali. Le seuil de sélection des individus représentatifs est fixé en analysant l'évolution de la distance quotidienne parcourue par rapport au taux de couverture de la journée en événement de communications. Nous montrons qu'au delà d'un taux de couverture de 60%, d'avantage d'information de communication n'implique pas une information de mobilité additionnelle. Ce seuil est donc suffisant pour que la description de la mobilité soit complète. En dessous de ce seuil, en revanche, il reste une sensibilité à l'activité de communication des utilisateurs qui est préjudiciable. Enfin, sur la base de cette calibration, nous effectuons une analyse de la distance parcourue sur une période historique. Les résultats démontrent notamment que les utilisateurs non réguliers participent à des tendances globales régulières et importantes, ce qui confirme que cette population doit être prise en compte dans l'étude de la mobilité vers une caractérisation complète du trafic et des émissions.

# Partie 3

Cette dernière partie de la thèse rassemble deux contributions importantes pour l'estimation des émissions atmosphériques à partir de données CDR:

- une méthode d'estimation des vitesses, nécessaire au calcul des facteurs d'émissions ;
- et une proposition d'intégration des briques méthodologiques développées dans cette thèse au sein d'une chaîne complète de modélisation et traitement de données.

### Chapitre 5: Estimation des vitesses moyennes de trafic à partir de données génériques sensibles aux activités de communication des utilisateurs.

Le chapitre 5 porte sur l'estimation de la vitesse moyenne du trafic, variable cruciale pour le calcul des facteurs d'émissions. On y présente un travail réalisé avant la réception des données du CDR. La méthode a donc été conçue pour traiter un format générique de données sensibles aux activités de communication des utilisateurs, appelées User Activity Dependent Positionning Data (données UADP par la suite). Ce chapitre est une version actualisée de l'article de journal :

Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, and Thamara Vieira da Rocha. Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths. Transportation Research Part C: Emerging Technologies, 129:103183, 2021.

Le format générique de données considéré dans ce chapitre englobe toutes les données massives de mobilité liées à l'utilisation des nouvelles technologies et dont la fréquence d'échantillonnage temporel dépend des comportements et activités de communication des utilisateurs. Il peut s'agir de données CDR ou d'autres types de données de téléphonie mobile, et notamment de données générées par des applications de réseaux sociaux comme Twitter ou Swarm.

De manière générale, la résolution spatiale et temporelle de ces données ne fournit pas des traces suffisamment riches pour estimer les vitesses individuelles et locales le long du trajet. Nous soulignons aussi que les données UADP impliquent, pour un trajet donné, des biais temporels sur les informations d'heure de départ et d'arrivée, et donc sur le temps de trajet total. Pourtant, étant donné les quantités massives dans lesquelles les données UADP sont disponibles, nous souhaitons démontrer qu'elles offrent une alternative viable aux données GPS pour estimer les vitesse moyenne du trafic en milieu urbain, notamment car leur grande quantité peut compenser leur faible résolution.

La méthode proposée exploite le découpage régional de la ville en régions. Celui-ci permet d'identifier des voyageurs traversant simultanément les mêmes zones et rencontrant *a priori* les mêmes conditions de circulation. Cette méthode repose sur une caractérisation simple des trajets extraits des données UADP, selon quatre propriétés: le chemin régional, l'heure de départ observée, l'heure d'arrivée observée et le temps de trajet observé qui découle des deux précédentes.

De manière périodique (toutes les 15 minutes), les voyageurs se déplaçant sur le même chemin régional sont regroupés, et la moyenne des temps de trajet observés sur le groupe est calculée. Nous montrons que sous certaines conditions, notamment de taille d'échantillon, connaître la moyenne du biais sur le temps de trajet permet d'estimer la moyenne du temps de trajet réel sur le chemin régional. La connaissance de ce biais temporel, spécifique à la fréquence de communication de la population considérée, permet l'estimation systématique des temps de parcours moyens sur le réseau routier à une échelle régionale.

Pour estimer les vitesses de trafic, nous suggérons d'analyser de façon conjointe les temps de parcours moyens sur les différents chemins régionaux, et de les considérer comme des contraintes sur les vitesses de trafic sous-jacentes. Selon ce principe, nous injectons les temps de parcours estimés un grand système linéaire. Ce système est construit de la sorte:

- les distances moyennes parcourues par région et chemin régional calculées dans le Chapitre 2 sont les coefficients linéaires du systèmes;
- l'inverse des vitesses moyennes régionales sont les inconnues;
- les temps de parcours moyen par chemin régional sont les constantes du second membre.

La construction et la résolution de ce système retourne un vecteur de vitesse moyenne qui caractérise l'ensemble des régions de la ville pour une période $t$. La reconstruction des dynamiques de vitesses sur une journée requière de construire et résoudre autant de systèmes que de périodes de temps dans la journée. La structure de cette méthode est particulièrement adaptée à toute entrée de données massives mais éparses, car elle nécessite très peu d'informations temporelles et d'itinéraires au niveau individuel, et tient compte du biais temporel inhérent aux déplacements extraits de ces données.

En raison du manque de données CDR réelles au moment où cette méthode a été développée, elle a été appliqué sur la ville de Lyon, en France, à des trajets artificiellement appauvris à partir de données GPS de véhicules. Pour conduire cet appauvrissement des données, nous développons un modèle de biais simple, qui n'a pas vocation à décrire le biais temporel réel des trajets extraits de données UADP, mais à simuler l'impact d'un modèle de biais simple mais réaliste sur l'estimation des vitesses dans le cadre de notre méthode. Cette démarche expérimentale permet d'abord d'utiliser les données GPS complètes pour estimer la vitesse réelle du trafic, référence à laquelle nous comparons nos résultat. Mais c'est aussi un moyen de tester différents scénarios d'appauvrissement des données, pour évaluer l'impact de l'agrégation spatiale ou des différents biais temporels sur la qualité des estimations de vitesse.

Les résultats principaux sont les suivants. Tout d'abord, nous démontrons qu'en l'absence de biais temporel, notre méthode reproduit de façon très satisfaisante les dynamiques de vitesses dans la majorité des régions du réseau, en dépit d'un niveau d'information déjà limité (chemin régional, heure d'arrivée, temps de trajet). Nous montrons ensuite que quelque soit le degré de bruitage des temps de parcours, la correction du biais en moyenne permet d'obtenir des estimations de vitesses satisfaisantes (bien que dégradées), à condition que la taille de l'échantillon soit suffisante. Enfin, nous évaluons l'impact des erreurs d'heure d'arrivée sur les résultats. A chaque étape, nous avons identifié des pistes méthodologiques pour limiter l'impact de l'appauvrissement des données. Nous avons en particulier identifié qu'une estimation dynamique plutôt que statique des distances moyennes de parcours permettait de réduire drastiquement les erreurs dans certaines régions. Notre étude confirme aussi que la qualité des résultats passe par l'intégration de larges échantillons de trajets. Dans le cadre de ce travail, nous avons été limités par l'origine des données (GPS), mais l'usage de données CDR garantit à priori l'accès à des données plus massives.

Une limite importante à l'application immédiate de cette méthode à des données CDR vient de la nécessité d'évaluer, pour ces données, le biais moyen des temps de parcours. Dans ce chapitre, nous avons proposé un modèle simpliste mettant en relation ce biais avec la distribution du temps inter-événement. L'objectif de ce modèle était de fournir un cadre méthodologique pour le sous-échantillonnage des données, et non pas de décrire avec précision le phénomène lui-même. Bien que la littérature se soit déjà penchée sur la caractérisation de la distribution des fréquences de communications dans données CDR et

réseaux sociaux, elle n'a à notre connaissance jamais exploré la question spécifique de la distribution du biais existant entre les temps de parcours observés pour les trajets extraits de ces données, et les temps de parcours réels. Pourtant, ces deux distributions sont intrinsèquement liées. Nous proposons dans la conclusion de ce chapitre et dans le suivant quelques pistes pour la caractérisation de ce biais.

### Chapitre 6: Chaîne de modélisation globale pour le calcul des émissions à partir de données CDR

Ce chapitre a pour objectif de proposer un cadre méthodologique global pour répondre à la problématique initiale de la thèse, à savoir le développement d'une méthode intégrée de traitement des données CDR pour l'analyse du trafic routier en milieu urbain, en vue de l'estimation des émissions atmosphériques qu'il génère.

La premère partie de ce chapitre présente la chaîne de modélisation et de traitement des données proposée. Cette chaîne de modélisation prend en entrée les données CDR, les données du réseau routier et des données complémentaires de recensement et d'enquêtes. Elle s'organise autour du processus d'estimation des vitesses et du traitement parallèle de deux catégories distinctes d'usagers (réguliers et irréguliers) avec des approches différentes. Cette distinction permet d'élargir considérablement le champ des usagers considérés et de traiter conjointement des profils de mobilité variés. La chaîne méthodologique se décompose en six étapes principales, qui correspondent pour la plupart aux différents chapitres de cette thèse :

1. Catégorisation des individus selon leurs profils de présence à partir de données longitudinales (observations sur plusieurs mois) et complément de cette catégorisation par l'estimation de facteurs de redressement pour chaque classe d'usagers.

2. Analyse du réseau routier pour identifier les chemins régionaux dominants et les longueurs régionales associées.

3. Reconstitution de la mobilité individuelle des usagers réguliers. Cette reconstruction se concentre sur deux aspects de leur mobilité : d'une part, l'enrichissement des chaînes d'activité à partir de l'identification des routines historiques de mobilité ; d'autre part, l'identification des chemins régionaux grâce aux résultats de l'analyse précédente.

4. Reconstruction de la mobilité des usagers non réguliers (locaux non réguliers et touristes) sur la base d'une approche collective, via une méthode hybride d'estimation des distances parcourues.

5. Estimation des vitesses de trafic par la fusion de grandes quantités de trajets dont l'information temporelle est biaisée.

6. Estimation des facteurs d'émissions en fonction de la vitesse de trafic et mise en relation avec les distances totales de déplacement pour une estimation globale des émissions.

Une figure importante de ce chapitre articule ces différentes briques et décrit les flux de données de l'une à l'autre. Elle précise aussi lesquels de ces traitements se basent sur une analyse temporelle de long terme. On y propose aussi un interfaçage possible de la chaîne de modélisation avec une méthode de détection des modes de transports.

Une deuxième partie de ce chapitre énumère les contributions scientifiques de cette chaîne de modélisation et des méthodes qu'elle intègre, et les rapporte aux objectifs de

recherche établis dans l'introduction. Ces contributions étant également résumées dans la conclusion, nous ne reprenons pas cette énumération ici.

Dans la dernière partie de ce chapitre, nous discutons trois des principales perspectives de recherche pour la mise en oeuvre de cette chaîne de modélisation dans un outils fonctionnel à destination de décideurs.

La première perspective de recherche, et la plus immédiate, est le développement de méthodes de caractérisation du biais temporel généré par la fréquence d'échantillonage variable des données CDR sur les informations de temps de trajets. Pour un territoire et une population donnée, une telle méthode permettra d'estimer la moyenne de ce biais, qui est indispensable au calcul des vitesses sur le territoire. Différentes solutions sont proposées, dont les coûts et dépendances en données supplémentaires sont discutés.

La seconde perspective d'amélioration de la chaîne de modélisation porte sur son interfaçage avec des méthodes de détection des modes de transports. L'articulation de notre travail avec de telles méthodes est nécessaire pour extraire de l'analyse les individus utilisant des modes doux, et pour associer les volume de trafic à des facteur d'émission cohérents selon le mode de motorisation. Nous présentons donc des suggestions de structure des flux et traitements de données, en portant une attention particulière à la question des temps de calculs et en questionnant la nécessité d'évaluer de façon systématique le mode de transport associé à un trajet. L'identification vraisemblable d'une régularité dans les choix modaux pourra notamment soutenir une implémentation légère de cet interfaçage.

Enfin, nous discutons de la dépendance de notre chaîne de modélisation à un certain nombre de phases d'apprentissage basées sur un historiques de données. La classification des utilisateurs selon différents profils de présence ou de régularité, et l'identification de leurs routines de mobilité correspondent à de telles phases d'apprentissage. Comme notre étude portait sur une période de temps limitée (3 mois), nous avons considéré que les résultats de ces analyses étaient constants. Des études de plus long terme nécessiterons de questionner la durée de validité des propriétés apprises sur la base de l'historique, et de les remettre à jour régulièrement. Nous introduisons la notion d'apprentissage continue, et discutons la possibilité de mettre en place des méthodes automatiques de détection d'anomalies pour déclencher de nouvelles phases d'apprentissage.

## Conclusion

Le sujet de cette thèse était le développement d'un cadre méthodologique pour l'estimation des émissions atmosphériques liées au trafic dans les zones urbaines à partir d'un type spécifique de données de téléphonie mobile, les données CDR. Ces données ont l'avantage d'être massives, disponibles et d'avoir une structure adaptée à l'étude de la mobilité individuelle, ce qui explique qu'elles aient fait l'objet de nombreuses études au cours de la dernière décennie. De par ces qualités, elles constituent une source de données alternative aux méthodes classiques d'analyse de la mobilité (enquêtes de mobilité ou détecteurs de boucles, par exemple).

Cependant, leur dispersion temporelle et leur faible résolution spatiale posent plusieurs problèmes lors de l'estimation de variables spécifiques du trafic. Premièrement, la résolution spatiale est une limite à la caractérisation des trajectoires fines et des distances de déplacement. Deuxièmement, les fréquences d'échantillonnage irrégulières constituent une limitation importante de l'estimation des déplacements origine-destination, des trajectoires et des flux. Troisièmement, ces caractéristiques empêchent l'estimation de temps de parcours précis et l'évaluation des vitesses de circulation individuelles. Enfin, le manque de statistiques supplémentaires sur la population étudiée limite l'interprétation de sa mobilité et oblige souvent à restreindre l'analyse à des profils d'utilisateurs spécifiques, comme les résidents.

Le cadre méthodologique élaboré dans cette thèse rassemble des méthodes issues de la littérature (et adaptées à notre problématique) et des solutions originales dans une chaîne de modélisation et de traitement des données cohérente. Les principales contributions de notre travail sont:

- La sélection d'échelles de travail mésoscopiques (régionales) et macroscopiques (urbaine), et l'adaptation d'une méthode de parcours automatique de graphe pour la caractérisation d'un réseau routier, de ses chemins dominants et de ses distances moyennes de déplacement régionales ;
- La classification des utilisateurs selon différents profils de mobilité et la formalisation de la mise à l'échelle de chacun des sous-échantillons de façon a garantir une reconstruction globale de la mobilité ;
- Le développement d'une méthode d'estimation des flux basée sur une approche couplant enrichissement des chaînes d'activités et reconstruction des trajectoires basée sur l'analyse d'un historique de mobilité ;
- La conception d'une méthode alternative pour estimer les distances totales parcourues sur le réseau par les utilisateurs irréguliers via une démarche collective et le concept de ratio de détour ;
- Le développement d'une approche innovante pour estimer les vitesses de circulation à partir de larges quantités de données biaisées caractérisées par une faible granularité spatiale et temporelle.
- L'articulation complète de ces différentes contributions dans une chaîne de modélisation permettant, à partir de données CDR et de données complémentaires limitées, d'estimer les variables de trafic nécessaire à l'estimation d'émissions atmosphériques.

Outres les contributions spécifiques des méthodes développées dans le manuscrit, le cadre méthodologique proposé constitue en soi une contribution essentielle de cette thèse puisqu'il fournit les clés pour estimer les émissions atmosphériques à partir des données CDR et de données complémentaires limitées. La fin de la conclusion de ce manuscrit dresse l'inventaire des perspectives d'amélioration et de recherche de ce travail. La poursuite de ces objectifs devrait permettre, à terme, l'exploitation complète de ce cadre sur un horizon de données long, d'abord sur la ville de Santiago de Cali, en Colombie, puis sur de nouvelles villes.

# Contents

# List of Figures

# List of Tables

# Introduction

## Background

The 21st century will have cities face substantial environmental challenges. Air quality and climate change are two of them. Fine particles and azote dioxide emissions frequently exceed alert levels fixed by the World Health Organisation, which exposes populations to significant sanitary risks and the cities and states to sanctions. Thus, cities aim to adopt efficient controlling tools for monitoring, notifying, and checking processes on air quality. Setting up those procedures contributes to the development of adapted solutions to pollution peaks. In the long run, it allows a better understanding of pollution phenomena, monitoring the impact of traffic and mobility management policies, and spreading the fight against climate change. Simultaneously, the urban population will not stop growing. This growth has already resulted in an intensification of urban mobility demand, the leading cause of urban pollutant emissions. Thus, it becomes urgent for urban decision-makers to manage the urban mobility requirements and the associated pollutant emission reduction in a joined and integrated way. The implementation of measures for sustainable traffic management is, therefore, both sanity and environmental necessity.

Although developed countries often possess such tools and acquire this best practice, such is not the case in developing ones, which usually face even worse population and urban growths and chronic traffic congestion situations. As the infrastructures are also lacking to support traditional monitoring techniques (loop detectors, air quality monitoring stations), there is a need for new and low-cost alternative methods and tools to monitor or evaluate public policies in transportation and urban quality.

In this context, the question of using diverse data sources to complete the sparsely available mobility information becomes a fundamental challenge. In this sense, this thesis questions the feasibility of estimating traffic-related emissions from cell phone data to assess the contribution of transport to pollutant emissions on a city scale. Although not directly intended for the observation of mobility, these data are a massive and insightful source of information about user behaviors. More specifically, the thesis focuses on estimating two main traffic variables, the traffic speed and the traffic volume, or total travel distance. These two variables are essential to derive the related traffic emissions. Call Detail Records (CDRs) have already been used to characterize key features of mobility patterns but when it comes to emissions, it becomes crucial to ensure completeness, *i.e.*, a characterization of mobility behaviors for all users and everywhere in the city. Temporal and spatial sparsity represent then a huge challenge that we aim to face in this thesis.

## State-of-the-art

Numerous traffic-related emission models have been developed until this day. These models integrate vehicles driving patterns to estimate emission factors, *i.e.*, the emitted quantity of monitored pollutant per unit of spent time, traveled distance or consumed fuel spent under the considered driving pattern [Smit et al., 2010]. The driving patterns are integrated

within this models in the form of different traffic variables, with varying temporal and spatial granularity. The granularity of these traffic variables can vary from very fine-grained, based on the characterization of the driving modes and cycles (with a periodicity of the order of a second), to a coarser level relying on the estimation of traffic situations or aggregated traffic variables (with a periodicity of the order of a minute or a dozen minutes) [Smit et al., 2010]. Since these emission models provide unit estimates of the emissions generated, estimating total emissions over time or over a population requires multiplying the emission factors by an estimate of the corresponding traffic volume, with a unit and scale consistent with the model. While the models based on the characterization of the driving cycles require precise inputs on the vehicle instantaneous dynamics, aggregated models call on more aggregated traffic variables such as total traveled distances or total traveled time under the focus traffic conditions.

Traditionally, the estimation of traffic-related emissions involves coupling these emission models with traffic models often based on traffic assignment models, either static or dynamic[Fallah Shorshani et al., 2015, Tsanakas et al., 2020]. These models allow to load a road network with a mobility demand and evaluate the related traffic states. These traffic states are further integrated into the emission model to derive the emission factors. [Vieira da Rocha et al., 2013] demonstrated the relevance of using microscopic dynamic traffic models for the estimation of fuel consumption, while [Samaras et al., 2017] coupled such a model with a fine-grained emission model to estimate the consumed fuel and $CO_2$ emissions in the city of Turin, Italy. We can also cite Smit et al. [2008] and Rodriguez-Rey et al. [2021] who estimated traffic-related emissions based on aggregated traffic and emission models respectively in Amsterdam, Netherlands, and Barcelona, Spain. The reader can refer to Tsanakas et al. [2020] for further examples of coupling between traffic and emission models.

With the advent of new technologies, especially in-car navigation systems, a new interest has emerged towards the characterization of pollution emissions and air quality levels from new data sources, such as GPS data. Nyhan et al. [2016], and later Kan et al. [2018] provide examples of coupling microscopic emission models with high resolution GPS data. At a coarser scale, some works aim to generate an integrated framework allowing to characterize specific traffic variables and related emissions. Shang et al. [2014], with GPS taxi data, proposed a model combining (1) an estimation of the average speed on each road segment from incomplete trajectory data ("context-aware matrix factorization" model), (2) a model to identify the volume of vehicles on each segment based on a Bayesian network, and finally (3) a calculation of energy consumption and pollutant emissions based on the results of the first two models. More recently, [Liu et al., 2018] proposed an emission method based on the fusion of taxi floating car data, license plate recognition data and geographical context data.

In the meantime, the significant democratization of cell phone use in the last two decades has opened up the field of relative data use in mobility analysis, and increased attention has been paid to the use of cell phone data to assess the environmental and health impact of mobility regarding air quality. Compared to floating car data, and especially taxi data, mobile phone data present some advantages. They display significant and still increasing penetration rates within the sensed population and more extensive spatial coverage (different modes or different paths). Several types of mobile phone data exist, characterized by different spatial and temporal granularities. We can, for instance, distinguish on one side the Location Based Services (LBS) and Location-Based Networking Services (LBNS), which correspond to data generated by smartphone apps, usually relying on the GPS, Wifi, and Bluetooth technologies, and generated in the background of the user activity. Because they often rely on GPS technology, these data display good spatial accuracy and high temporal sampling rates. Another type of data often used in mobility

analysis corresponds to signaling data. These data, generated by the communication network, allow geo-locating mobile phone users at any time, at various spatial scales, either base station or location area (group of base stations). Finally, Call Detail Records (CDRs) collect the user-generated communication events, such as calls, texts, or data connections at the base station scale. Because its collection rate depends on users' communication behaviors, this data is generally less intrusive than signaling data from a privacy point of view. Their spatial scale also makes them more protective than LBNS data. They are therefore most readily shared by communication operators than the two other types. This thesis focuses mainly on this specific type of data, but the following state-of-the-art has been drawn up keeping an eye as well on other types of data, which in some aspects remain similar.

Most advanced researches in the field of environmental evaluation based on this kind of data have looked into the estimation and prediction of the air quality and pollution level from features extracted from mobile sensing data. These approaches directly relate the features extracted from the mobile sensing data with pollutant levels observed at monitoring stations. The correlations between those two sets of data are explored and used to estimate and forecast the pollutant levels based on the mobile sensing data. In Zheng et al. [2013, 2015] for example, the air quality levels are either estimated in real-time or forecasted with machine-learning techniques. These methods rely on local and regional air quality data (both historical and real-time), weather data (historical or forecasts), traffic-related features (distribution of traffic speeds), mobility-related features (number of people entering and leaving the grid cell, based on the pickup and drop-in data), road networks and points of interest properties. Though these methods extract traffic and mobility features from the mobile phone data, they are only used first as training, then as input data, without contributing to an explicit representation of the global mobility in the studied city. Those models seem efficient to infer real-time or forecast future air quality levels from historical or present air quality-related features (mobility, weather) and the local or regional (and once again historical or current) trends in air quality levels. There is no need for transportation mode detection or scaling up from the mobile sensed user to the global population. However, such an approach requires deploying many fine-grained air quality sensors on a large scale, which can have high costs. Moreover, those traditional machine learning-based solutions do not integrate a representation of traffic to explain the air quality factors they characterize or predict. Such an approach prevents isolating a specific cause (*e.g.* traffic) from the numerous possible confounding factors (other human activities like industries, weather, air quality historicity) and to related it to its specific consequence (traffic-related emissions). Hence, it prevents determining the causes of the observed trends in air quality and evaluating how (transportation) policies will effectively impact air quality. In order to accurately evaluate the effects of such policies on mobility, traffic, and hence traffic-related emissions, it becomes necessary to implement a more stepwise approach, such as presented in the following part. This direct relating of communication data with air quality is an interesting approach, however, the air quality measures are sensitive to other confusing factors, such as economic and industrial activity or weather.

But the specific development of emission estimation methods based on such technologies, and especially CDRs, is scarcer than with GPS data. Indeed, the intrinsic data limitations make them mostly inappropriate for traffic characterization, although they were shown to be suitable for general mobility analysis. Among emission-oriented mobile-phone related works, we can cite the framework proposed by Li et al. [2016] that specifically involves estimating pollutant emissions by grid mapping CDR data. One of the specificities of the framework is that a connectivity matrix is constructed based on the road network to describe how well two adjacent cells are connected, depending on the number of direct

connections (roads) that link the first cell to the other. Transition costs between non-contiguous cells are deduced from this connectivity matrix. When the grid mapping of a trajectory results in detected positions in non-contiguous cells, then the most likely path between them is supposed to be the shortest path between those cells computed based on those transition costs. After computing the users' traces, the total traveled distance is calculated within each cell as the product of the number of CDR users crossing the cell within a given time window and the constant length of the cell. Though the framework proposed is the first step into an innovative approach, it presents some limits. Firstly, the transition cost matrix relies on the grid connections but not on the capacity of the links or their speed limitation: the shortest path computed can therefore be imprecise. Moreover, the authors do not consider the existence of multiple possible paths between two non-contiguous cells and therefore do not render the multiplicity of possibility behind the observed trajectory. Last but not least, the total travel distance is essentially based on the constant grid length, which does not render the potentially very distributed length of trips inside one cell. However, this is an attractive, easy-to-implement basic approach.

The development of an integrated framework that is rigorous on this issue raises the question of evaluating a certain number of variables necessary for applying emission models. As previously stated, traffic speed and volume are essential to traffic variables for estimating the emissions relating to the circulation of an urban population. However, the estimation of these variables from mobile phone and CDR data is far from being trivial. In this paragraph, we present the state-of-the-art of these methods. CDR-based studies have mostly focused on the analysis of land use [Becker et al., 2011, Furno et al., 2017] and mobility recurrent features [Gonzalez et al., 2008, Jiang et al., 2017]. Among these topics, the question of the estimation of origin-destination matrices has been widely explored [Iqbal et al., 2014, Alexander et al., 2015, Toole et al., 2015]. However, most of these studies present two major limitations.

First, they base their analysis on a specific subset of the population, which are the residents of the study area. Although the focus on these users is legitimate for the purpose of reconstructing origin-destination matrices of local populations, this approach is restrictive when it comes to characterizing traffic, since it ignores a significant part of the population which, although not resident in the area, contributes to road traffic and thus to air emissions. This includes individual profiles such as visitors, external commuters or people in transit for instance.

Second, most of these studies did not account for the biases introduced in the matrices by the data temporal sparsity. This is problematic, considering that the time gaps may hide visited locations, and result in erroneous matrices and origin-destination flows estimations [Zhao et al., 2021]. With synthetic CDR data, Hoteit et al. [2017] and Chen et al. [2018] evidenced the impact of the temporal gaps on classical mobility statistics. In Chen et al. [2019], the authors addressed the question of the data sparsity by proposing a method to enrich the daily trajectories for hourly time slot. They further re-evaluate well known results of the literature in the light of the new completed data, such as mobility laws and predictability. However, the impact of this completion on traffic-related variables, such as origin matrices or traveled distances, was not further investigated. Among the few exceptions Zhao et al. [2021] estimated that 10% of the visited locations might be missing in a CDR data, when proposing a method to infer whether an observed CDR-derived trip was hiding a latent activity.

Not only the temporal sparsity of the data may affect the estimation of origin-destination matrices and flows, but it may also impact the detection of trajectories, and therefore of the local traffic volumes. To the best of our knowledge, this question, and the aim of estimating path flows from CDR data has barely been addressed before, although it was addressed with other types of mobile phone data, like location-based networking services

[Paipuri et al., 2020]. Lwin et al. [2018] proposed a method for estimating hourly link flow from CDR data, using a shortest path analysis at the network scale. However, one could argue that this method is not the most adapted to the spatial resolution of CDR data, and may result in overestimate flows of very specific routes. Without specifically aiming at estimating the flows, a range of the literature has focused on reconstructing the route taken by mobile phone users. However, such methods have been much more largely developed with signaling data (more frequent) [Asgari et al., 2016, Bonnetain et al., 2019] than CDRs Forghani et al. [2020]. Therefore, estimating path flow from CDR data remains an open question.

Other strategic traffic variables are also at stake. The map-matching challenges raise the underlying question of the estimation of individual travel distance. Most works on the subject have focused on estimating long travel distances at nation-wide or international scales Nilbe et al. [2014]. At an urban scale, besides cell-to-cell shortest paths [Lwin et al., 2018] or euclidean distances [Li et al., 2016], the literature cruelly lacks research on this subject. When it comes to estimating traffic speeds, the literature is even poorer. Based on synthetic signaling data, Ou et al. [2011] proposed a method for estimating the traffic speed on an isolated freeway segment, while Derrmann et al. [2017] and later Paipuri et al. [2020] used respectively real signaling and location-based networking data to estimate macroscopic fundamental diagrams. To the best of our knowledge, no comparable work has been done on CDR data, and no alternative was provided to estimate traffic speeds.

## Research objectives and contributions

### Emission model and scale requirements

The definition of the research objectives emerges from a back and forth reflection between the constraints imposed by the data and the objectives at the end of the emission modeling chain. On the one hand, the ambition to work with CDR data imposes a minimal spatial granularity corresponding to the cellular network density. This granularity, but also the low temporal sampling rates of the data, prevent the reconstruction of microscopic driving patterns and corresponding traffic variables like high-resolution link flows. On the other hand, the size of the targeted perimeters (urban scale) raises the question of the complexity of the methods implemented and suggests selecting an emission model adapted to this scale, such as the aggregated emission models like COPERT [Ntziachristos et al., 2009]. The popularity of such models at institutional levels (they are used for local to national emission inventories) is an additional element in favor of these models, since the methods developed in this thesis are intended to be integrated into a decision support tool for institutions and local authorities [Ntziachristos et al., 2009].

Therefore, this thesis aims at providing the appropriate traffic inputs for a coupling with aggregated speed-based models such as COPERT. Formally, for a pollutant $k$, the corresponding emissions $E^k$ (in g) are calculated as:

$$E^k = TTD \cdot F^k(V) \tag{1}$$

where $TTD$ stands for Total Travel Distance (in km), $V$ for the traffic speed (in $\mathrm{km\,h^{-1}}$) and $F^k(V)$ corresponds to the emission factor (in $\mathrm{g\,km^{-1}}$). In practice, emissions can be further distinguished by vehicle class, *i.e.*, specific characteristics of vehicle and motorization. The specific scientific issues concerning the characterization of vehicle fleets and of the modal shares are out of the scope of this work, and, in the context of the Green City Big Data (GCBD) project, processed by other project partners.

Therefore, the traffic variables targeted in this thesis are of two kinds. First, the traffic speed $V$ allows estimating emission factors, *i.e.*, unit amount of pollutants emitted. Second,

estimating the volume of traffic, or total distance traveled under these speed conditions allows to estimate the overall emissions.

### Research objectives

Many CDR- or mobile-phone-data-based research have focused on reconstructing traffic variables at the road network level. In this thesis, we investigate the problem of estimating these variables at an intermediate regional scale, which we believe is more appropriate for the data than the road network scale. A first objective is therefore to define this scale, whose minimum granularity would be imposed by the spatial resolution of the CDR data, and then to estimate traffic variables adapted to this scale. Further on, let $R$ be the regional partitioning of the network and let $r$ be any region of this partitioning. At this scale, and for each time slot $t$, the variables we target are the total travel distance $TTD_r(t)$ and the dynamic mean spatial speed $V_r(t)$. The specific temporal and spatial characteristics of CDR data, that prevent the identification of users' trajectories, make this a non-trivial objective.

To estimate these two target traffic variables, one of the main research objective will be the data completeness and mobility enrichment, at the individual and group level. At the individual level, we will ask ourselves how to reconstruct the mobility of a data user, in terms of activity chains and trajectories. We will also question the possibility of reconstructing the individual mobility of all users. At the group level, we will be interested in the question of the selection of the individuals to be included in the analysis, and we will explore how to proceed to the most accurate expansion of the sample, and how to estimate the mobility when it cannot be reconstructed at an individual level. These question will be essential to estimate traffic volumes.

Besides, we will try to address the question of the speed estimation in a new light. From the observation that the irregular temporal sampling rates of the data and their variability from one user to the other prevent the estimation of clear travel times, we will look into developing a data-fusion method to evaluate traffic speeds from trips data without inferring individual speeds.

Finally, our last research objective is to related these specific works in a global framework aiming at estimating the traffic variables necessary for the estimation of emissions.

## Contributions

This section summarizes the contributions of this thesis.

- The first main contribution corresponds to the original introduction of a user classification method with differentiated scaling and mobility reconstruction methods. This coupling aims at contributing to a global and exhaustive estimation of the traffic volume by considering broader sections of the population than the literature.
  - We propose a classification of users based on:
    - Their presence pattern in the area, to separate resident, commuters, and visitors.
    - Their mobility regularity, to separate regular from irregular mobile users.
  - We extend the scaling methods from the literature, that focus on resident users, by formalizing scaling methods for two additional categories of users, external commuters (commuters living outside of the monitored area) and visitors.
- To support these mobility reconstruction methods, we adapt a method from the literature to estimate the prevailing regional paths and traveled distances. The first contributions focus on the travel distance estimation from CDR data. Rather than trying to estimate the distances right away from the CDR data, which are spatially imprecise, we propose to rely on an MFD-based literature work to estimate such average distance from systematic network analysis. We propose two low-level modifications to implement it on a large scale, as in our case study.
- We propose differentiated mobility reconstruction methods according to the user classes.
  - For regular users, we propose both activity-chains and trajectory reconstruction methods.
    - Historical mobility routines are identified using a clustering method. They are integrated within a heuristical approach and support the enrichment of light daily activity chains.
    - The sparse trajectories are reconstructed based on the previous prevailing paths identification.
    - We propose an adaptive path flow distribution estimation method to map-match the overall trip set.
  - For irregular users, we propose a large scale total traveled distance estimation.
- The last contribution of this thesis concerns the development of a method to estimate traffic speeds from noisy mobile sensing data. We propose a methodological data fusion workflow to estimate the regional traffic speed from user-activity-dependent positioning data. We test it in a controlled environment with artificially biased trips. We show that even though the data quality does not allow to estimate individual traffic speed, processing a large number of individual trips can compensate for the data temporal biases. Although the method has not been tested yet on real CDR data, this will be the subject of future work.
- The final contribution of this thesis is the organization of these methods into a large framework for estimating the global traffic-related emissions in an urban context.

## Thesis Outline

This thesis follows the following outline. It is divided into three parts.

The first part introduces the data, the case study used for the most significant part of this thesis, and the preliminary data processing that support the implementation of the overall framework proposed in this thesis. In particular, Chapter 1 presents the territory of the Greater Cali area and the quality and specific features of the CDR data that was made available for this research work. Chapter 2 introduces the major and usual limitations of the traffic volume estimation from CDR-data: 1. the reduction of the population to a favorable (resident) population, 2. the estimation of the path and the traveled distances. This chapter presents two preliminary analyses of the CDR data and the road network that will support the implementation of the overall framework and allow to deal with the limitations aforementioned.

The second part of the thesis focuses on the traffic volume reconstruction for two different categories of the population.

Chapter 3 focuses on the share of the population considered to be regular. This includes local users, such as residents and commuters, that display common mobility characteristics (based on an entropy measure). This chapter presents a heuristic for enriching those users' daily activity chains based on constructing individual mobility profiles from the identification of spatial routines. This chapter further presents our path flow estimation method, based on the overall path database enhancement, through the map-matching of a subset of trips and the distribution of flows accordingly.

Chapter 4 instead focuses on the non-regular users, a category which includes the non-regular resident and commuters, but also the non-local ones, such as visitors. Considering that their mobility might not be representative (and therefore legitimate to be upscaled) at the path flow level, we directly focus on the total travel distances of the population. We adopt an aggregated distance estimation method based on the detour ration function.

The third and last part of the thesis focuses on emission-oriented contributions. In Chapter 5, we present an innovative traffic speed estimation method based on synthetic user activity-dependant data. This chapter corresponds to an article published in a peer-reviewed journal. In Chapter 6, we present the overall integrated framework and its related We identify the most significant perspectives and remaining works for it to be fully implementable.

Finally, we conclude this work in the Conclusion section.

# List of publications and communications

## Peer-reviewed journal papers

- Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, and Thamara Vieira da Rocha. Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths. *Transportation Research Part C: Emerging Technologies*, 129:103183, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103183. URL `https://www.sciencedirect.com/science/article/pii/S0968090X21001996`

- S.F.A. Batista, Manon Seppecher, and Ludovic Leclercq. Identification and characterizing of the prevailing paths on a urban network for mfd-based applications. *Transportation Research Part C: Emerging Technologies*, 127:102953, 2021b. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102953. URL https://www.sciencedirect.com/science/article/pii/S0968090X20308500

## Peer-reviewed conference proceedings

- Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri. Dynamic Estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths, *hEART 9th Symposium*, Lyon, France, 2020.

## International conference presentation

- Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri. Dynamic estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths, *Transportation Research Board 99th Annual Meeting*, Washington DC, USA, 2020.

## Other communications

- Manon Seppecher, Dynamic estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths, Scientific Seminar of LICIT, Lyon, France, 2020.

- Manon Seppecher, Dynamic estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths, *ERC MAGnUM annual meeting*, Corrençon-en-Vercors, France, 2019.

- Manon Seppecher, Dynamic estimation of Urban Zonal Speed from Mobile Sensing Data and Macroscopic Paths, *IFSTTAR-COSYS Annual Seminar* La Baule, France, 2019.

- Manon Seppecher, GCBD Project: Using mobile phone data for estimating urban traffic and related air emissions, *GAUNAL Scientific Seminar*, UNAL, Medellin, Colombia, 2019.

- Manon Seppecher, Using mobile phone data for traffic estimation and pollutant emissions evaluation, *Workshops Mobility & Big Data and Air Quality*, Centro Mario Molina & AFD, Mexico City, Mexico, 2019.

# Part I

# Framework initialization: case study and data fusion

The first part of this thesis aims to lay the foundations for comprehensive mobility analysis of an urban territory.

Understanding the territory and its population is necessary to reach such an objective. The characterization of the data available for the analysis is equally required. These two points are brought together in Chapter 1. This chapter also proposes a regional partitioning of the base station network based on the data features. It supports the mobility estimation methods developed in the following chapters.

The reconstruction of mobility from cell phone data raises several issues. These issues are of two kinds. First, in the absence of individual socio-economic and demographic data on users, the problem is to characterize the sample and upscale it accurately. Second, the fragmented nature of the cell phone data, both spatially and temporally, raises the question of thoroughly estimating the flows and distances traveled by the study population. Chapter 2 attempts to provide answers to these two problems. It is structured around two methods that aim to characterize the sample and the territory studied in greater detail. The first method classifies the sampled individuals according to their presence in the area. It will support the development of adapted expansion methods. The second concerns the road network studied. It provides constants to characterize the typical paths, and average distances traveled on this network at a regional level. These data provide a basis for interpreting user mobility.

---

Outline



---

# Chapter 1

# Case Study

In this chapter, we introduce the studied perimeter, along with the data accessible, and their impact on the spatial resolutions of our analysis.

## 1.1 Studied perimeter

In this thesis, we focus on the city of Santiago de Cali, Colombia, and its greater area.

Santiago de Cali is the third most populated city in Colombia, behind Bogota and Medellin. The municipality is located in Valle del Cauca department, in the South West of the country. It is divided in 22 *comunas* or urban districts, and 15 *corregimientos* or rural districts. The work developed in this thesis deals specifically with the characterization of traffic within the city bounds, therefore precisely within the 22 urban districts. Yet, this characterization requires considering a larger area which will provide insightful information about the mobility within the city.

The municipality of Cali is strongly related to two neighboring municipalities, Yumbo in the north and Jamundi in the south. Even though these areas are out of the geographic scope of our study, the population of these municipalities, along with the inhabitants of the rural areas of Cali, contribute to the traffic and mobility within the Cali's center. Therefore, the population of these areas will be considered as well in order to characterize the commuting mobility patterns towards and from the city. These administrative divisions are illustrated in Figure 1.1.

Cali's urban area is 123 square kilometers large, while the area of the overall municipality is 569 square kilometers. According to the latest national census [DANE], 2 172 527 people leave in the urban area of Cali, while 55 115 leave in the rural districts. Table 1.1 provides the complete population and surface indicators for the municipalities of Yumbo, Jamundi and Cali at aggregated levels.

| | Total | Municipality | | | | |
|---|---|---|---|---|---|---|
| | | Jamundi | Yumbo | Cali | | |
| | | | | Total | Urban Area | Rural Area |
| Population (mil.) | 2.72 | 0.13 | 0.13 | 2.46 | 2.43 | 0.03 |
| Area ($km^2$) | 1434 | 632 | 234 | 569 | 123 | 446 |
| # of BS | 440 | 26 | 41 | 371 | 339 | 32 |
| # of BS per $km^2$ | 0.36 | 0.04 | 0.18 | 0.65 | 2.79 | 0.7 |

Table 1.1: Comparison of the perimeter properties in term of geography and available data

At a finer scale, Figure 1.2a displays the distribution of the population density in the municipality of Cali at the comuna scale. This is put into perspective with a map of

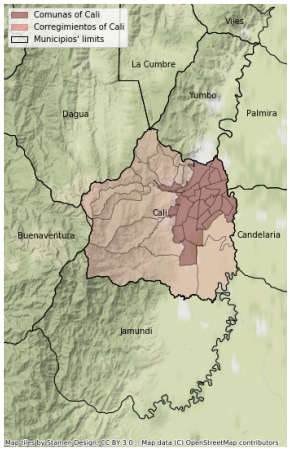
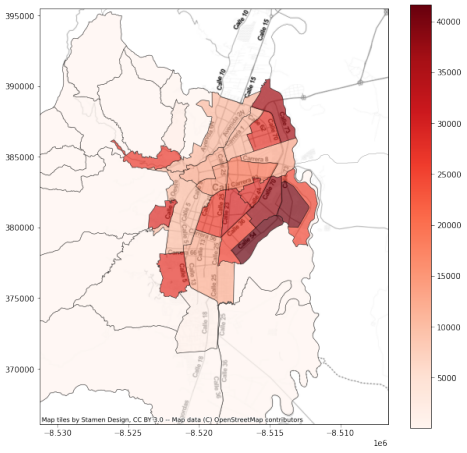
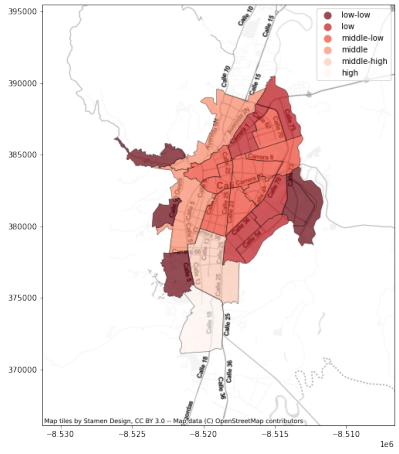



Figure 1.1: Valle del Cauca administrative division




(a) Population density (inhabitants per km$^2$) in Cali municipality




(b) Median socio-economic stratification per urban district

Figure 1.2: Density and socio-economic characteristics of the territory

the economic stratification of the city at the same scale (Figure 1.2b). We can observe a correlation between high density and low median socio-economic stratifications. The peripheral urban district, both in the east and west side of the city are characterized by high densities and very low stratification. In the inner city, the eastern districts appear to be both more populated and with lower median stratification. Instead, the west districts and southern districts appear to be less populated and with higher median stratification.

## 1.2 Available CDR Data

The mobile phone data used in this thesis are Call Detail Records (CDR) data, provided by the American data provider CLARO. CDR data are a specific type of mobile phone data, which are passively generated by registered users when communicating, and stored by the data provider for billing or network management purposes. These data record the spatio-temporal information of communication events, and more specifically the starting timestamp of the communication event and the location of the base station that processed it, along with the anonymized user id information. Communication events include emitting or receiving a call, exchanging texts, or browsing data. Complementary properties are also registered, such as the technology used (*e.g.*, 3G, 4G), or whether the event was incoming or outcoming.

| User ID | Base station | timestamp | event type | technology | emission/reception |
|---|---|---|---|---|---|
| A | $BS_1$ | 09:10 | sms | 3G | incoming |
| A | $BS_1$ | 09:20 | sms | 3G | outgoing |
| A | $BS_1$ | 17:40 | call | 3G | outgoing |
| A | $BS_2$ | 21:30 | data | 4G | incoming |

Table 1.2: Raw data structure

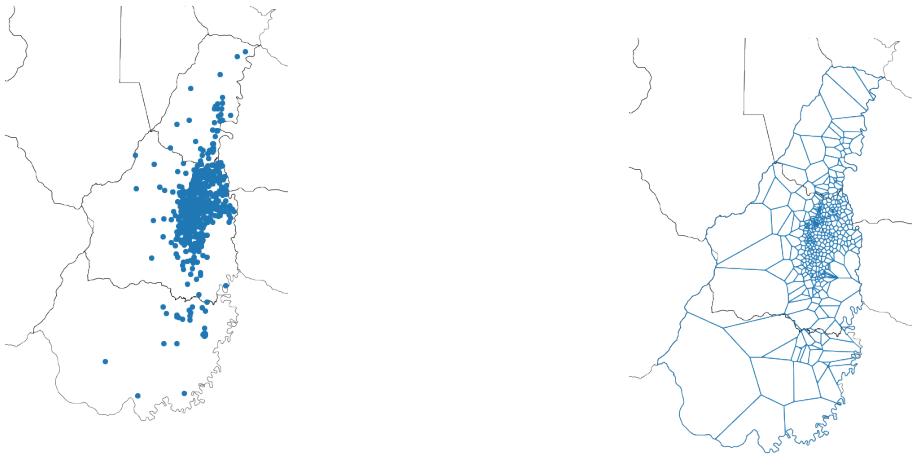
| User ID | Base station | First timestamp | Last timestamp | # of events |
|---|---|---|---|---|
| A | $BS_1$ | 09:10 | 17:40 | 3 |
| A | $BS_2$ | 21:30 | 21:30 | 1 |

Table 1.3: Compressed data structure

While the structure described above is the basic CDR data structure, the data format was modified before transfer, for both privacy protection and compression of size of the dataset. In particular, the data was compressed so as to keep only the mobility information. The format transformation operated by CLARO was conducted as follows. Users' consecutive communication events were aggregated into a communication sequence entry when occurring at the same base station. The resulting communication sequences were characterized by the user's id, the base station location, the timestamps of the first and last event of the sequence, and the number of events observed during the sequence. The differences between the original and compressed datasets are illustrated in Tables 1.2 and 1.3.

This aggregation process of several communication events into a single communication sequence presents the advantage of significantly reducing the data size. It also restricts the user information transmitted to the minimal mobility information carried by the data. However, it erases precious information on the users' communication rhythms, since all the intermediate communication timestamps are lost. This closes the door to analyzing communication rhythms within static sequences and the possibility of inferring missed movements in-between the static events.

The data transmitted by CLARO covers a period of 23 months, starting from January 2020 to November 2021. The first three months were exploited in this thesis, during which a total of 2 707 012 unique user IDs were observed over the covered area. In Valle del Cauca, the data covers the municipality of Santiago de Cali, along with the neighboring municipalities of Yumbo and Jamundi. The data shared by CLARO also provided coverage of Medellin, but this city is out of the scope of this thesis, and its data was barely used. The following information and statistics are provided for the data observations in Valle del Cauca only.



(a) Distribution of the 440 base stations over the municipalities of Cali, Jamundi and Yumbo

(b) Voronoi tesselation of the territory based on the base station positions

Figure 1.3: Cali's base station network

A total number of 440 base stations are covering Cali, Yumbo and Jamundi, with an uneven density. The base station network is denser in the city center than in the city surroundings, where approximately 70% of the base stations are located. Therefore, the network density there is of 2.79 base stations per square km, while it is of 0.16 base stations per square km outside of Cali.

Figure 1.3 illustrates the inhomogeneous distribution of the base stations over the territory. In Figure 1.3a, we observe the higher concentration of the base stations over the urban area of Cali. Although less dense, the network remains relatively continuous and well supplied to the north towards Yumbo. Towards the south, on the other hand, the antennas become rarer and except for a higher concentration around the urban center of Jamundi, it is in this area that their concentration is lowest. The general lower network density in the western part of the area is explained by the lower population density and the higher relief.

Figure 1.3b represents the Voronoi's tesselation of the BS network. The Voronoi's tesselation is the unique partitioning of the perimeter into as many polygons as base stations, in which the polygon associated to a base station corresponds to all the positions that are closer to this base station than to any other. This tesselation is classically used in the literature to identify the geographical area covered by a tower, and assume the possible positions of a user signaling from that base station. Using this tesselation to infer the users positions comes back to assuming that each users' event is processed by the base station that is the closest. This hypothesis might not be systematically verified. In fact, the antenna that will process an events might be further away, depending of the network density, the base station altitude or the network load. However, it is a reasonable hypothesis to which we will stick for this entire thesis. Therefore, this partitioning of the network represents the maximal resolution we can expect from the available mobile phone data. This resolution will be further discussed in the following section.

Along with the geographical indicators discussed in Section 1.1, Table 1.1 provides a comparison of the BS network density across the different zones of the perimeter.

## 1.3 Spatial resolution definition

### 1.3.1 Challenges and selected approach

After displaying the Voronoi tesselation of the base station network, it is clear that the varying density of the base stations across the area will result in varying uncertainties in the positioning of sampled users. While dense areas with small Voronoi polygons will be associated with low positioning uncertainties, these uncertainties will rise significantly in the less dense areas of the network, where the Voronoi cells can cover several square kilometers.

Along with the heterogeneous communication network, the heterogeneous urbanization and population of the area may also result in uneven user samples sizes between base stations. Those points raise the question of the comparability and of the statistical representativeness of the mobility and traffic indicators that we aim to derive and monitor from the analysis of the data. Therefore, we propose in this section aggregation methods to generate, on the basis of the Voronoi tesselation of the network, less granulated partitionings of the area with enhanced area and population homogeneity across regions.

It is also worth noting that the Voronoi partitioning of the network obviously does not respect the administrative partitioning presented in Section 1.1. We have indicated in this section our willingness to derive mobility and traffic variables at the Cali city level only, but our intention to do so by taking into account the contributions of sampled users from neighboring municipalities. Therefore, we must consider transposing this administrative distinction at the base station level and in the cell aggregation process.

For this reason, we suggest in a first step dividing the base station network into two large sub-areas, the city of Cali versus its surrounding rural districts and municipalities. This process is described in Section 1.3.2. This first partitioning of the network will provide a way to label antennas with urban and non-urban information.

In a second step, described in Section 1.3.3, we will spatially aggregate base stations of each of the two sub-networks into regions of approximately homogeneous size and numbered of sampled users.

### 1.3.2 Urban and rural sub-areas definition

To start with, we divide the base station network into two concentric sub-areas. Those sub-areas are supposed to split the network between urban base stations covering Cali's city and the outer areas. As the tessellation defined by the Voronoi's polygons of the BS network do not match the administrative boundaries of the city, it is necessary to tolerate that the urban perimeter defined from the base stations does not strictly match the city area. We propose to associate to the city of Cali all the base stations that are located within a buffer of 700 m around the administrative border of the city. Compared to selecting only the base stations located within the city administrative bounds, this buffer allows to integrate really close base stations whose Voronoi's polygons are partially covering the city area.

The resulting perimeter is displayed in Figure 1.4. Thereafter, we will call $C_0$ the inner region and $C_1$ the outer area.

We will see in the following section how this division of the base station network into two sub-areas is essential prior to the base station aggregation step. But from the mobility analysis perspective, this division will also provide a quick and light criteria to categorize users as urban or rural residents and compare their behaviors. More details will be provided on this point in Section 2.

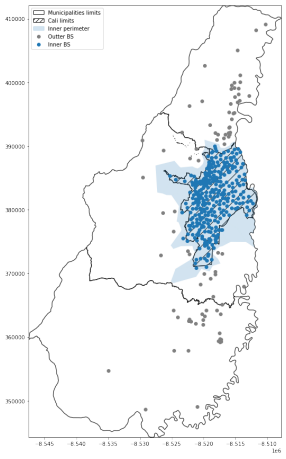



Figure 1.4: Partitioning of the network in two sub-networks

### 1.3.3 Sub-networks aggregation

From the base station aggregation perspective, the division of the base station network in two sub-areas is essential as it provides the geographic limits that the aggregation should respect. Indeed, one could consider gathering base station cells together on area and population criteria without consideration of whether the cells in question are mainly rural or urban cells. However, this would result in regions with urban and rural cells mixed together, while we want our final zoning to be as consistent as possible with the existing administrative zoning. Therefore, instead of running a single aggregation process over all cells, we propose running it twice, once based on the cells of $C_0$, once based on the cells of $C_1$.

Our proposed method requires that not only the area of each base station cell be known, but also that the sample of users residing in that cell be identified. The matching of users with a residence cell is described in Chapter 2. Here, let us simply assume that this information is already known. The method also requires a minimal sample size and area objectives $s_{obj}$ and $a_{obj}$ to be set. We will fix $s_{obj}$ to 400 residents and explore various values of $a_{obj}$.

The pseudo-code of the overall algorithm is described in Algorithm 1. In a first step, the algorithm aggregates cells together based on the number of residents matched with that cell. The purpose of this phase is to ensure that all regions have a corresponding population sub-sample large enough and a reliable statistical representativity of the information derived from this sample. In a second step, the algorithm aggregates regions together based on their area. This phase aims at ensuring that regions have similar areas.

For one criterion or the other, the aggregation process is the same and follows a simple greedy algorithm. The initial regional network $R$ strictly corresponds to the Voronoi tessellation of the network $V$. If any region $r$ have the considered characteristic (sample size $r.s$ or area $r.a$) below the pre-defined minimal objective ($s_{obj}$ and $a_{obj}$), we select among such regions the region $r_0$ that has the lowest value for that characteristic. That region $r_0$ is merged with its neighbors $r_1$ whose centroid (or BS location in the first iteration) is closest to its own. The centroid position is re-calculated, and the sample size and surface of the new region are updated as the sum of the features of the merged regions. The process is iterated until all regions meet with the sample size or area objectives.

In our analysis, the population criterion is set to $s_{obj} = 400$ individuals. The impact of this first aggregation is illustrated in Figure 1.5. In Figure 1.5a, the base station network is represented with level of colors according to the size of the corresponding resident sample. Hashed cells represent cells with sampled population below the fixed threshold. Figure 1.5b

**Algorithm 1** Base station network aggregation

**Require:**

Voronoi partitioning of the base station network $V$.

$V = \{c \in V\}$ where $c$ is a Voronoi cell.

$v.g$ is the cell geometry, $v.c$ is the base station location, $v.a$ is the cell area and $v.s$ is the size of the related user sample.

$s_{obj}$: the minimal sample size objective

$a_{obj}$: the minimal area objective

**Ensure:**

New regional partitioning $R$ as aggregation of the BS cells and meeting the population and area objectives

Initialization:

$R \leftarrow V$

▷ 1: Aggregation based on population criterion

**while** $\exists r \in R | r.s < s_{obj}$ **do**

  $r_0 \leftarrow (r$ with smallest $r.s)$

  $r_1 \leftarrow$ closest region from $r_0$ ▷ dist. calculated between $r_0.c$ and $r_1.c$

  ▷ Merge regions

  $r_1.g \leftarrow r_0.g + r_1.g$

  $r_1.a \leftarrow r_0.a + r_1.a$

  $r_1.c \leftarrow centroid(r_1.g)$

  $r_1.p \leftarrow r_0.p + r_1.p$

  delete$(r_0)$

**end while**

▷ 2: Aggregation based on area criterion

**while** $\exists r \in R | r.a < a_{obj}$ **do**

  $r_0 \leftarrow (r$ with smallest $r.a)$

  $r_1 \leftarrow$ closest region from $r_0$ ▷ dist. calculated between $r_0.c$ and $r_1.c$

  ▷ Merge regions

  $r_1.g \leftarrow r_0.g + r_1.g$

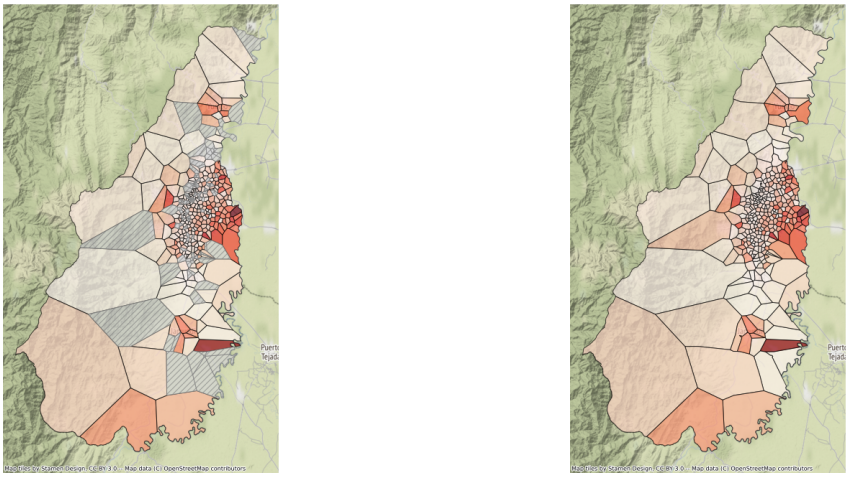
  $r_1.a \leftarrow r_0.a + r_1.a$

  $r_1.c \leftarrow centroid(r_1.g)$

  $r_1.p \leftarrow r_0.p + r_1.p$

  delete$(r_0)$

**end while**



(a) (b)

Figure 1.5: Aggregation of cells based on the population threshold. a) Raw Voronoi partitioning of the BS network. Cells with hashes represent cells with a number of resident detected below the threshold of 400 individuals. b) Resulting partitioning after the aggregation of Voronoi cells.

displays the new network once cells have been iteratively aggregated together. We observe that the cells merged during this phase are mainly rural cells. Despite the large areas covered, these cells are associated with smaller sub-samples than those in the urban area due to lower population densities but perhaps also lower use of communication technologies.

On the contrary, the aggregation based on the surface criterion has a much larger impact in $Z_0$ than in $Z_1$. As the zoning within this sub-area will define the resolution of the outputs of the subsequent methods, we consider several values for the parameter $a_{obj}$ and various resolution configurations. More precisely, we select two different area thresholds that result in two zoning with a very different number of regions. The selected area thresholds are 100 ha and 900 ha. They result respectively in zonings of 149 and 49 regions, in which respectively 93 and 16 regions are urban regions. The resulting zonings are displayed in Figure 1.6, and referred to as $\mathscr{R}_1$ and $\mathscr{R}_2$. Figure 1.7 summarizes the distribution of zones surfaces in each cases. In the urban sub-area, the first aggregation level correspond to aggregating together 3 to 4 cells. It therefore corresponds to a significant yet still granulated partitioning, and will present the advantage of keeping a high spatial resolution of the results. Instead, the second aggregation scale with a partitioning of 16 urban regions more drastically scales up the BS network and reduces its complexity. Although far from the standard Voronoi partitioning, this zoning presents characteristics (in number of regions and area of zones in particular) comparable to those of the partitioning frequently adopted in the MFD literature, on which our methods partially relies. In the following of this thesis, we will explore the result of the modeling chain on the regional partitioning $\mathscr{R}_1$. These resolution will especially define the scales at which the traffic variables will be observed. This range of different spatial resolutions will provide insight into the sensitivity of our methods to the complexity of the regional network and the amount of regional data.


Figure 1.6: Plots of the two regional partitioning $\mathscr{R}_1$ and $\mathscr{R}_2$

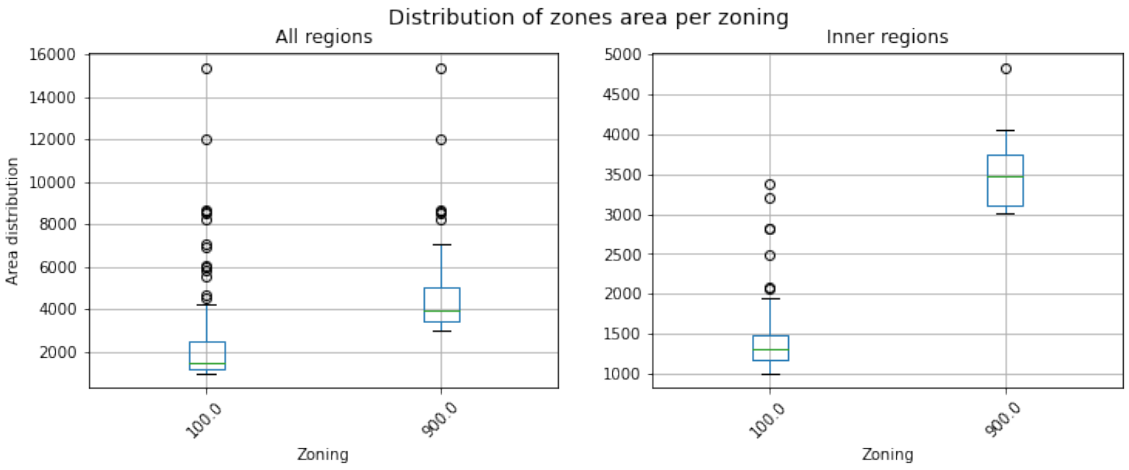



Figure 1.7: Boxplots of region areas for each parameter, for the overall perimeter or the urban one

# Chapter 2

# Population and network preliminary analysis

The characterization of the total traveled distance $TTD_r$ within a region $r$ requires to estimate both the travelers flow $q_r$ and their average trip lengths $\bar{L}_r$ in $r$ as:

$$TTD_r = q_r \cdot \bar{L}_r \tag{2.1}$$

This chapter aims at providing methodological background for the accurate estimation of those two traffic variables.

For estimating the volume, one needs to take into consideration all urban users and characterize as precisely as possible their mobility patterns. Identifying from CDR data various urban behaviors and up-scaling them according to the magnitude of the sub-population they represent within the real urban population is a critical point of our methodology. In Section 2.1, we highlight the challenges raised by the scaling issue and formalize a methodological solution. This solution relies on a user categorization method based on the identification of different mobility profiles. This categorization is supported by an extensive population analysis.

When based on cell phone data, given the data low sampling rates, estimating regional volumes also raises the issue of detecting users within all the regions crossed along their path, and not just the regions where they generate CDR data. Moreover, the specific temporal sampling rates, but also the low spatial resolution, are at stake for estimating the distances traveled by individuals, required for the calculation of the average regional trip length ($\bar{L}_{r,p}$). We propose to rely on a literature work from the MFD-theory domain to face this two-folded challenge. This method is independent of any trip sample and relies on a systematic road network analysis to characterize at the regional scale both the prevailing paths between regional origins and destinations, and the average lengths along those paths. Section 2.2 introduces this work and presents the adaptations we conducted for applying it to our case study.

## 2.1 Sample up-scaling: challenges and approach

### 2.1.1 Introduction and related works

Working with user samples to estimate mobility and traffic variables such as flow requires up-scaling the sampled-based results to characterize the overall population. This process goes through expanding and weighting user sub-samples according to their potential under or over-representation compared to reality. It usually relies on the identification of subgroups of users sharing similar characteristics, *e.g.*, residence area, or socioeconomic

| | Sample spatial coverage | Sample duration | Filtered users |
|---|---|---|---|
| Nanni et al. [2014] | Ivory coast | Several samples of 2 weeks, with re-sampling in-between | Users with too little activity from their home or work locations |
| Çolak et al. [2015] | Metropolitan areas of 2 cities | At least 2 months | Users with too little activity from their home or work locations |
| Alexander et al. [2015] | Boston's metropolitan area | 2 months with partial re-indexation | Users with less than 1 visit to house location per week on average |
| Toole et al. [2015] | Metropolitan areas of 5 cities | At least 3 weeks | Users that make fewer than 2.5 trips per day |
| Lwin et al. [2018] | Nation-wide (Myanmar) | 1 week | · |

Table 2.1: Comparison of the OD matrices studies

features. The size of these local sub-samples is then compared to the corresponding census populations and up-scaled accordingly.

The individual socioeconomic data being rarely available in CDR-based literature, the grouping criteria is the users' residence area. As communication data providers usually do not share the users' home location either, the field studies have recourse to methods called Home Detection Algorithms (HDA) to identify the census block of residence of the sampled users. These algorithms belong to a more extensive literature field that identifies meaningful places from individual mobile phone traces, such as home, work, or leisure locations. HDA measure various geolocated activity indicators to evaluate the significance of each visited place in the mobility pattern of a user and generally infer the most important place to be the home location. These indicators include, for instance, the number of days or the total time spent at the considered location, or the number of events generated from that place [Nanni et al., 2014]. The reader can refer to Rojas et al. [January 2016] and Vanhoof et al. [2018] for surveys and benchmark on this subject. These indicators may also be measured on limited time windows, such as nights or weekends, when users are assumed to be mostly at home [Alexander et al., 2015, Çolak et al., 2015, Osorio-Arjona and García-Palomares, 2019]. When the spatial coverage of the data is nation-wide (such as in Lwin et al. [2018] in Myanmar, Vanhoof et al. [2018] in France), any user can be assigned a home location, and hence a scaling factor. Indeed, the spatial coverage being maximal, the home locations of the sample (national) users are necessarily within this coverage, and the authors can safely assume that it correspond to the most significant place detected. However, such favorable spatial coverage is rarely available in CDR data research projects. Most of the samples made available for research in the literature have limited spatial coverage at the city or metropolitan area scale (see Table 2.1). In the latter case, assigning a residence location to users living outside the covered area (commuting, visiting, or transiting users) becomes a challenge. Either no home location is assigned to them (*e.g.*, if they are not in town during the night-time home detection period), or it is erroneously assigned. Therefore, the mobility analysis in the CDR-based literature focuses on the users who are assumed to be residents and excludes others. It is a regular limitation of the works estimating Origin-Destination (OD) matrices, despite substantial literature [Nanni et al., 2014, Çolak et al., 2015, Alexander et al., 2015, Toole et al., 2015]. Table 2.1 gathers some of the most significant works of the domain and transcribes the filters implemented for this purpose. Although these methods are consistent with estimating the OD matrices of resident populations, they ignore a significant part of the individuals who do not live in the area but who contribute, with specific mobility patterns, to the area traffic and air pollution. Those individuals include users such as commuters or visitors, for instance. This latter category has been the subject of specific mobility studies [Sikder et al., 2016].

However, those studies are often carried thanks to roaming mobile phone data that is peculiar to foreign visitors [Nilbe et al., 2014]. They also focus on the large regional or national scale visiting patterns rather than the urban ones [Vanhoof et al., 2017]. Lastly, these works draw conclusions from the sample's observations without adjusting and up-scaling the samples to extend their conclusions to overall visitors populations.

Given our objective of evaluating the total traffic volumes, extending the CDR-based mobility analysis framework from resident users to a more extensive range of urban behaviors represents a critical milestone of our work. Reaching this goal requires identifying and defining the non-resident complementary user categories and classifying the surplus sample according to those. It also involves formalizing for each new category an up-scaling approach that does not rely on home location knowledge. In essence, the works listed in Table 2.1 already proceed with such categorization, as they separate presumed residents from the rest of the sample. This filtering process ensures that the individuals included in the analysis are sufficiently present or active to be considered as residents. Users not meeting these criteria are simply discarded. This filter is therefore a marginal element of these works in the literature, in the sense that it simply makes it possible to pre-define the scope of the study in terms of the sample studied. Its impact on the results does not appear to be explored nor considered to be a key issue. However, once we consider including a more significant share of individuals in the analysis, this criterion for characterizing resident and non-resident users becomes a central analysis element. The OD-matrix-centered literature has mostly neglected this question so far. To the best of our knowledge, no study integrates varied mobility profiles within an extensive traffic variable estimation framework.

However, some works have focused explicitly on classifying CDR data users, yet without pulling the thread of this classification towards a traffic analysis. For instance, Furletti et al. [2012] propose a categorization method based on a step-wise improving process. First, the authors propose a top-down categorization process that distinguishes users' profiles based on a set of pre-defined and restrictive rules or constraints defining three categories: residents, commuters, and transit users. Second, a bottom-up unsupervised learning process refines the categorization. It classifies the remaining users based on their presence patterns. This refining process is applied to identify the visitors out of the mobile users. The authors further develop this work in Furletti et al. [2013], where they aggregate user data into call profiles. First, those individual call profiles are clustered using the same unsupervised learning method, and then the resulting groups are gathered using K-means. Here, the authors focus on three different user categories: the residents, the commuters, and the visitors, and discard the transit users. Those works are further continued in Gabrielli et al. [2015], with a simplification of the user classification: the authors propose applying K-means directly on the individual call profiles before labeling the resulting clusters using expert-based archetypes. This work results in four mobility profiles: residents, dynamic residents (going out of the area for work), commuters, and visitors. In Mamei and Colonna [2018], the authors adopt a method close to the one developed in Furletti et al. [2012]. A subset of sampled users, respecting stringent rules, are tagged with different mobility profiles (resident, tourist, commuter, transit, and excursionists). Then different classifying methods are tested, and the C4.5 algorithm, which returns satisfactory accuracy results, is selected for further analysis. Finally, Thuillier et al. [2018] present an interesting alternative approach, which starts with the labeling of the different days of observation of users instead of labeling users themselves. The authors infer different daily behaviors from the daily user observation span. They include transit and commuting behaviors and different resident behaviors. The authors also assign an *absent* label when users are missing during the considered day and a *weekend* label for observations occurring on Saturdays and Sundays. This daily labeling supports a weekly vectorized representation of users. The set of weekly patterns of the population is then clustered to identify representative weekly

| | Categorization method | Sample duration |
|---|---|---|
| Furletti et al. [2012] | Labeling with stringent rules and clustering (SOM) | 1 month |
| Furletti et al. [2013] | Double classification: users then user groups | 1 month |
| Gabrielli et al. [2015] | K-means on Individual Call Profile vector and expert labeling | 5 weeks |
| Mamei and Colonna [2018] | Labeling with stringent rules and supervised classification (C4.5) | 2 non-consecutive months |
| Thuillier et al. [2018] | Weekly vector construction from day labeling + clustering of weekly user patterns | 3 weeks |

Table 2.2: Categorization methods developed in the literature

| | Residents | Dynamic residents | Commuters | In Transit | Tourists | Excursionist |
|---|---|---|---|---|---|---|
| Furletti et al. [2012] | ✓ | · | ✓ | ✓ | ✓ | · |
| Furletti et al. [2013] | ✓ | · | ✓ | · | ✓ | · |
| Gabrielli et al. [2015] | ✓ | ✓ | ✓ | · | ✓ | · |
| Mamei and Colonna [2018] | ✓ | · | ✓ | ✓ | ✓ | ✓ |
| Thuillier et al. [2018] | ✓ | ✓ | ✓ | ✓ | ✓ | · |

Table 2.3: User profiles considered in literature

patterns.

This literature review evidences the large range of urban mobility profiles one can consider: residents, dynamic residents, commuters, users in transit, tourist, or excursionist (see Table 2.3 for a comparison). However, it neglects the user profiles up-scaling questions. This limits the reach of these works since the classified volumes and flows that could be derived from them require scaling up. On one side, several user profiles raise the question of the user's origin, and therefore of the identification of the group it belongs with. On the other side, obtaining reference data about these sub-population sizes to support the up-scaling is very difficult, if not impossible for some users categories. While national and regional censuses provide baselines of the residents, urban and touristic travel surveys often lack a standardized format and methodology that could easily provide the key numbers that a rigorous expansion of the other user categories requires. These surveys also often focus on specific users types, loosing track of the rest of the population. This is for instance the case of touristic surveys, that focus on some touristic behaviors but miss other visiting users.

Despite these limitations, we propose categorizing users according to three different mobility profiles (resident, commuter, and visitor) and formalizing scaling approaches adapted to each category. These scaling approaches are based on the inventory of data carried out at the scale of Colombia and the metropolitan area of Santiago de Cali. They are therefore necessarily relative to our case study. Nevertheless, we believe that this work can lay the foundation for broader and more systematic analyses. In the process of both classification and scaling, we stress the importance of working over a sufficiently long period. It is interesting to note that, despite significant methodological contributions, neither the OD-matrix-focused works nor the profile-categorization studies have addressed the issue of the sensitivity to the duration of the study. However, it seems to us that this is a critical element that may impact the separation or categorization of individuals, especially when

dealing with the shortest time frames (one week to one month). In those latter cases, long-stay visitors can be confused with residents, while residents may go unnoticed if they are on vacation or simply not very active. Tables 2.1 and 2.2 report on the temporal dimension of each literature work. We consider a large historical period (two months) to support our user categorization, and discuss in further details why this sample duration is also at stake for the up-scaling process in a dedicated paragraph.

Section 2.1.2 goes through the definition of the objective categories and the methodological approach we follow for categorizing and up-scaling the corresponding individuals. Section 2.1.3 exposes the various results of the application of this methodology to our case study. Lastly, Section 2.1.4 discusses the method and its limitation and proposes improvement perspectives.

### 2.1.2 Methodology

#### 2.1.2.1 Objective categories

Extending the target population of CDR-based analysis calls for the selection of complementary user profiles to be considered beside the resident users. Despite the large range of urban behaviors, we focus on two additional mobility profiles to minimize the number of user categories as each type considered raises the challenge of population baseline for sample extension, but also requires adapted mobility processing methods. Commuters and visitors are these two additional categories.

Below, we propose a definition for each objective category.

1. Residents $R$: individuals living in the area covered by the antennas;

2. Commuters $C$: individuals that live outside of the area but enter it on a frequent basis;

3. Visitors $V$: individuals that mainly live and work outside of the area, but may visit the studied territory, either for touristic reasons with a dense stay, or from time to time with shorter stays.

It should be mentioned here that to this first level of classification, a second level of classification will be added based on the regularity of mobility. This double classification is more fully discussed in the introduction to Part II, where it supports the implementation of differentiated mobility reconstruction methods.

#### 2.1.2.2 A simple binning approach

The literature review suggests several methods for user categorization. Our objective being to implement a light-weighted and easy to implement classification method, we first opted for the approach proposed by Gabrielli et al. [2015]. However, the data structure in our case was found inadequate for the application of a K-means clustering: the data displays an high continuity in the users' behaviors and individual call profiles, and no natural cluster could be identified. As a consequence, the K-means clustering results in an inappropriate and apparent arbitrary division of the data.

Rather than applying such arbitrary limits on the individual categories, we use a simple binning approach, where the limits of the clusters are calibrated based on our observations of the individual behaviors. Although this approach also results in drastic data partitioning, it presents the advantage compared to the K-Means algorithm to set the user division transparently. We consider this binning implementation as a first, easy-to-implement user categorization method. In a context where the user categorization is also tough to validate due to the lack of ground truth data, it also allows calibrating the binning threshold according to macroscopic constant variables.

Based on the definitions selected previously, we consider as significant discriminating features the following ones:

- $f_{day}$: the number of days of observation in the area;

- $f_{weekDay}$: the number of week days of observation in the area;

- $f_{night}$: the number of nights with observation in the area;

- $f_{maxStay}$: the shortest stay (in number of consecutive days) observed over the historical period.

| | | Binning rules | | Role |
|---|---|---|---|---|
| | | $f_{night} > t_{high}$ | | Present at night |
| Residents | or | $f_{night} > t_{low}$ and $f_{day} > t_{high}$ | or | present at day (w/ softer night condition) |
| | or | $f_{maxStay} > t_{high}$ | or | has at least a long stay |
| Commuters | | $f_{weekDay} > t_{high}$ and not a resident | | Present at day |
| Visitors | | User is not a resident nor a commuter. | | Other users |

Table 2.4: Binning rules

While $f_{day}$ and $f_{weekDay}$ will isolate local users (residents and commuters) from visitors, $f_{night}$ will allow to separate within local users the commuters from the residents. Introducing $f_{maxStay}$ allows enriching the distinction between residents and commuters, commuters being expected not to appear in the area for too many consecutive days, in relation with week-ends for instance. The definition of $f_{day}$, $f_{weekDay}$ and $f_{night}$ requires the definition of corresponding time windows. We partition the day in four time-windows: night (8p.m.- 7a.m.), early morning (7a.m.- 9a.m.), restricted day (9a.m.- 6p.m.) and late afternoon (6p.m - 8p.m). The two-hours long early morning and late afternoon time window are considered as transition time windows during which the users' positions can be varied through the time window and from day to day. We prefer not to infer the user profile based on this period and will focus on the night and day time windows, during which users are assumed to be more static, to estimate their user profiles. The day-based features are calculated on the restricted day window, and the night-based one during the night period.

Based on the categories definitions provided above and the presence features selected here, we explicit a set of binning rules, summarized in Table 2.4. Those rules rely on only two threshold parameters, $t_{high}$ and $t_{low}$. The threshold $t_{high}$ aims at setting a high presence threshold that guaranty the users meeting this criteria are local users: they are present enough in the area, at night or day, to be considered either resident or commuters. The threshold $t_{low}$'s purpose is to extend the resident category too users who are sufficiently observed at day to be considered locals, and sufficiently observed at night to be considered residents instead of commuters. These categories are roughly summarized in Figure 2.1.

#### 2.1.2.3 Threshold calibration

Calibrating the thresholds $t_{low}$ and $t_{high}$ is a necessary step to apply the binning rules onto the user dataset. We propose a calibration approach supported by the relatively large spatial coverage of the data. On one side, this spatial coverage (Santiago de Cali, Jamundi and Yumbo municipalities) implies that within the residents $R_1$ of $Z_1$ (Yumbo and Jamundi), a share of the population commutes towards $Z_0$ (Cali). Let us call those commuters $C_1$. On the other side, the local mobility survey and census data provide references in the resident and commuting population sizes. Therefore, we propose deriving from those two reference ratios to which the classified user data can be related:

$$r_1^{ref} = \frac{|R_1^{ref}|}{|R_0^{ref}|} \tag{2.2}$$

$$r_2^{ref} = \frac{|C_1^{ref}|}{|R_1^{ref}|} \tag{2.3}$$

where $R_0^{ref}$ and $R_1^{ref}$ respectively refer to the reference number of residents in $Z_0$ and $Z_1$. Similarly, $C_0^{ref}$ refers to the reference population of residents living in $Z_1$ and commuting

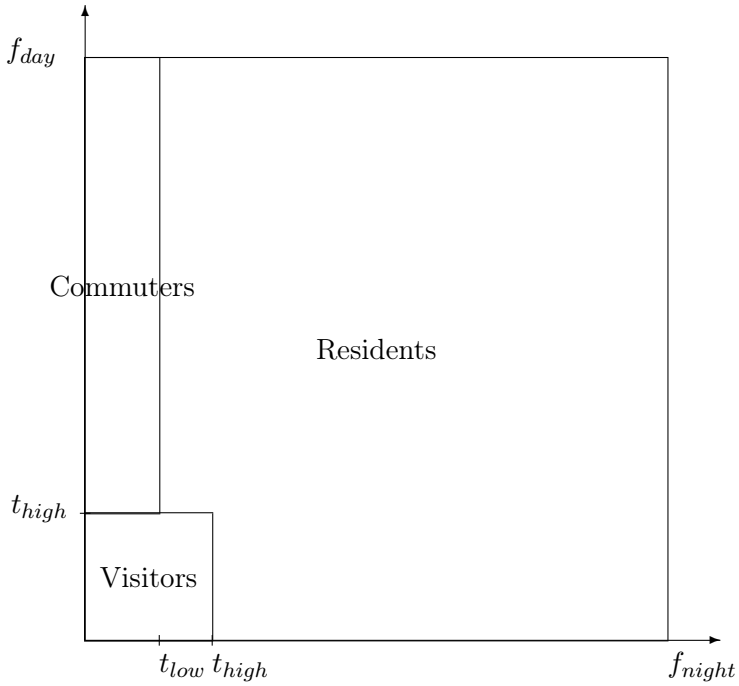



Figure 2.1: User classification diagram in the $(f_{night}, f_{day})$ plan

towards $Z_0$. The ratio $r_1^{ref}$ describes the relation between the suburban and city population sizes, while ratio $r_2^{ref}$ characterizes the share of the suburban population that commutes to the city. We consider these ratios as targets that the categorization of individuals should aim at, assuming that:

1. the penetration rates of the mobile technology between $R_0$ and $R_1$ are identical;
2. the penetration rates of the mobile technology between $C_1$ and $R_1$ are identical.

These assumptions allow us to consider that these ratios $r_1^{ref}$ and $r_2^{ref}$ are not affected by sampling bias and can be considered valid at the sample level. Although these assumptions might not be true when the two areas compared belong to two very different geographic contexts, such as rural versus urban areas, and might also not hold at a fine geographic scale since communication habits also vary with socio-economic characteristics, we consider here a relatively aggregated scale, with two regions well urbanized, which supports these assumptions.

The national census [DANE] and Cali's mobility survey [Metro Cali, 2015], both conducted in 2015, provide us with following ratios:

$$r_1^{ref} = 12\% \text{ and } r_2^{ref} = 33\% \tag{2.4}$$

Within this calibration framework, we propose to go through the parameter space and select the couple $(t_{low}, t_{high})$ that provides the $r_1$ and $r_2$ ratios the closest to the references. In practice, we adopt the following processing chain.

In a first step, we apply a HDA to assign all users a *potential home location*. In practice, the potential home location is assigned to the base station where the users were observed to spend the most nights at. The location where a user spends their night is assumed to be the place where they generate the most data from, during the night time-window. The presence features $f_{day}$, $f_{weekDay}$, $f_{night}$, $f_{maxStay}$ are also extracted from the historical data at two different scales: $Z_0 \cup Z_1$ and $Z_0$.
Let $f^{Z_0 \cup Z_1} = (f_{day}^{Z_0 \cup Z_1}, f_{night}^{Z_0 \cup Z_1}, f_{maxStay}^{Z_0 \cup Z_1})$.
Let $f^{Z_0} = (f_{day}^{Z_0}, f_{night}^{Z_0}, f_{maxStay}^{Z_0})$.

Then, we iterative explore the parameters space. For each parameter couple $(t_{low}, t_{high})$, we classify users according to $f^{Z_0 \cup Z_1}$ following the rules set established above. We retrieve from the categorization the resident users and use each user home location to distinguish between $R_0$ and $R_1$, and using $f^{Z_0}$ features, we identify users of $C_1$ out of $R_1$. We can compute $r_1$ and $r_2$, and so we iterate in order to find the best threshold couple.

Figure 2.2 illustrates this overall workflow. Results are presented in Section 2.1.3

#### 2.1.2.4 Scaling strategies

The up-scaling of a sub-sample $s$ requires two elements:

- The size of this sub-sample $|s|$.
- The identification of the population $P_s$ it represents, and of its cardinality $|P_s|$.

From this, the sample scaling factor $f_s$ is defined as:

$$f_s = \frac{|P_s|}{|s|} \tag{2.5}$$

Therefore, a rigorous scaling factor calculation implies consistency between the selected sub-sample and the reference population.

This consistency should first be spatial, of course. This criterion supports the well-explored scaling of residents in CDR-based studies. Once matched with a home base-station, the resident is further distributed over the different census blocks it intersects, following area and population weighting method formalized by Bachir et al. [2017]. Then, for a given census block $b$, the scaling factor $f_{R,b}$ of the residents in $b$ is:

$$f_{R,b} = \frac{|R_b^{ref}|}{|R_b|} \tag{2.6}$$

where $R_b^{ref}$ is the block population according to the census data, and $R_b$ is the sample of residents assigned to $b$.

However, a second consistency criterion is the temporal span, and it is barely considered in the literature. Indeed, not only a longer time span allows more robust user categorizations, but it only contributes to the temporal consistency of the expansion procedure. Working with a sample of great historical depth allows to include larger number of users, and therefore to be temporally consistent with, in the case of residents, the census data, which is usually calculated over a year. It allows to relate the population observed during the overall historical period with the census data, while respecting the possible day-to-day or week-to-week variations that occur within the sample period. This temporal consistency issue is even more important when considering dynamic populations, with important presence variations or turn-over, like commuting users or visitors. We discuss this question in Section 2.1.3.2. But for commuters and visitors, the spatial consistency issue is also at stake, since their place of origin is out of the monitored area, and unknown. Based on an extensive administrative data inventory, we identify two references for up-scaling the commuting and the visiting users.

For commuters, the only reference document available is a general mobility report generated from a 2015 mobility survey [Metro Cali, 2015]. This report provide the number of travel between the urban centers of two municipalities (Palmira and Candelaria) of the greater area of Cali and the city. Considering that they correspond to trips between

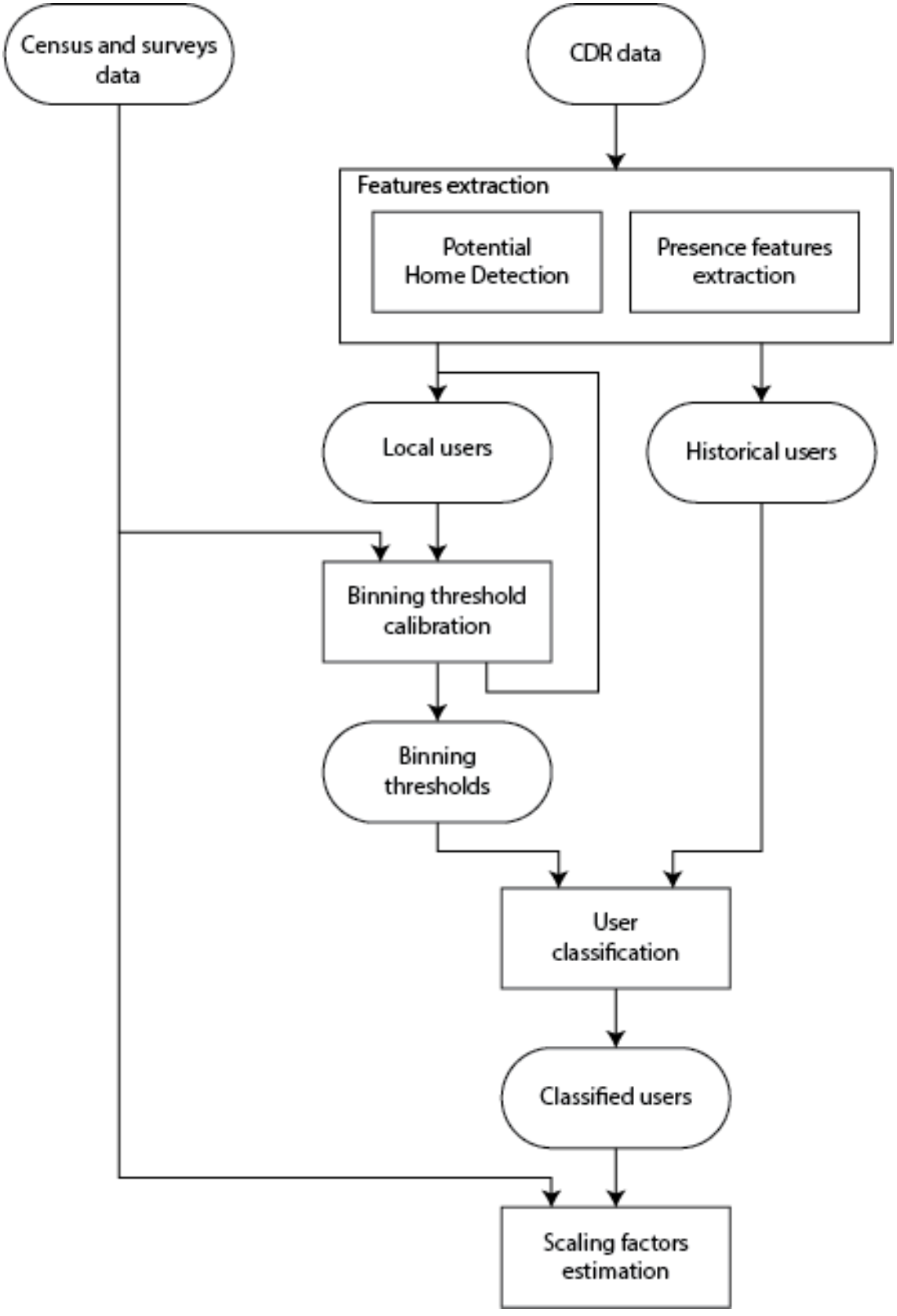



Figure 2.2: Data processing for user classification

spaced-out municipalities, we assume that a trip is equivalent to a commuter. This is a very restrictive data, since it characterizes the population of a limited territory among the regions that provide travelers towards Cali. However, as it is our only reference, we consider it as our baseline for the commuting behaviors. This means that we assume that every commuting user comes from those two municipalities centers. This sample must therefore fit the magnitude of that baseline population as reported by the selected mobility survey.

$$f_C = \frac{|C^{ref}|}{|C|} \tag{2.7}$$

When it comes to visitors, they correspond to a very varied population. They can include nearby residents who come to the city only a few days a month for shopping or business reasons, or tourists that stay in the city for several days, as part of a trip in which Cali is or is not the main destination. However, the reports on this population are often generated by tourist and immigration services and focus on this latter tourist profile. In the case of Cali, these reports [Secretaría de Turismo de Cali, 2019] also focus specifically on users whose primary destination is Cali, thus eliminating visitors to other Colombian cities. For lack of more precise data, we will rely on this reference for the rest of our analysis.

Visitor users, and especially touristic users, are also characterized by import temporal dynamics and a high population turnover, on the contrary to residents and commuters users that are more stable populations because they are locals. The influx of visitors especially shows strong seasonality [Baum, 1999]. Therefore, a yearly estimate of the number of visitors $|V^{ref}_{year}|$ cannot be fully satisfactory to estimate individual scaling factors. In our case, no monthly data could be found about the national visiting dynamics in Cali, although a yearly baseline for 2019 $|V^{ref,national}_{year}|$ was identified [Secretaría de Turismo de Cali, 2019]. However, such monthly data was available for international visitors. Let $|V^{ref,international}_{month}|$ be this monthly reference. Assuming that the international and national visitors follow the same trends, we inferred the monthly national visitors as follows:

$$|V^{ref,national}_{month}| = \frac{|V^{ref,international}_{month}|}{|V^{ref,international}_{year}|} \cdot |V^{ref,national}_{year}| \tag{2.8}$$

Processing to a monthly up-scaling of visitor users requires to know the total number of sampled visitors in the month $|V_{month}|$. When dealing with day-by-day data, as we intend to do, this number may be unknown. Estimating a daily baseline from the monthly reference would be problematic for two reasons. First, because visitors may stay several days, therefore the daily number of visitors is different from the monthly number of visitors divided by the number of days in the month. The second reason is that some events may generate strong but punctual attractions in the city: scaling to a constant population of daily visitors would crush the daily variations of the visitor population. Therefore, we suggest learning the average individual scaling factor from a preliminary historical period of at least a month.

$$f^{hist}_V = \frac{|V^{ref}_{hist}|}{|V_{hist}|} \tag{2.9}$$

Then, we can proceed with the day-to-day data analysis and expansion for the following data using the learned scaling factor $f^{hist}_V$. After processing the full data of the current month, its scaling factor can be corrected by computing the ratio between the expected visitors population size (from the touristic surveys) $|V^{ref}_{current}|$ and the effective number of visitors detected $|V_{current}|$ during the month:

$$f_V^{current} = \frac{|V_{current}^{ref}|}{|V_{current}|} \tag{2.10}$$

The historical scaling factor $f_V^{hist}$ can also be further re-calibrated.

Before, proceeding with the results in the following section, some of the assumptions made in this paragraph call for a discussion. Indeed, we have made strong assumptions about the sample population to relate them to the available baseline numbers. On one side, we propose to consider that all the observed commuters are coming from the centers of Candelaria and Palmira, while they could come from more rural or remote areas. On the other side, we suggest to relate all the visitors observed to tourists making their main stay in the city of Cali, while they could be non-tourist visitors or traveling from or towards another significant destination. These two assumptions have several consequences. In practice, they mean that we might actually be proceeding with a population reduction instead of expansion, in term of scope or population size. By scope, we mean that a certain number of users will be reduced to a profile that is not necessarily their own. By size, we mean that the actual number of users observed by be larger than the category reference, resulting in scaling factor smaller than 1. This means that the resulting mobility and traffic variables will only aim to represent the reference populations only (residents, nearby commuters and touristic visitors), and still not the overall urban population.

### 2.1.3 Results

This section presents the various results of this categorization and up-scaling chain. We present the calibration of the categorization method, the conclusion drawn from this categorization, then the results of the up scaling process.

#### 2.1.3.1 User categorization

The first steps of the categorization method are the extraction of the presence features and the detection of a potential home base station. Based on the time-windows previously defined, we measure the number of days and nights users are observed in $Z_0 \cup Z_1$. Figure 2.3 illustrates the users distribution in the $(f_{night}, f_{day})$ plan. Based on the night time window, we also infer the potential home base station of each users. The results of this potential home detection method are presented in Figure 2.4.

Proceeding with the calibration of the two thresholds $t_{low}$ and $t_{high}$, we make them vary to identify the couple that matches the best the local specific features of the reference population. We make $t_{low}$ vary between 2 and 10, and $t_{high}$ vary between $t_{low}$ and 20. The resulting ratios evolution, along with the objective ratio values, are displayed in Figure 2.5. Figure 2.5a displays the evolution of the macroscopic ratio $r_1$ with $t_{low}$ and $t_{high}$. Figure 2.5b displays the evolution of the macroscopic ratio $r_2$ with $t_{low}$ and $t_{high}$. While we observe that $r_2$ is very sensitive to $t_{high}$, and can find a value for which results match the objective $r_2$ value (33%), the evolution of $r_1$ stabilizes above a plateau around 18%. It suggests that we overestimate the population of $Z_1$ compared to $Z_0$. However, we suspect an under-estimation of the census data. Indeed, while the finer-grained census data are population projections based on the 2015 national census [Armitage Cadavid et al., 2019], a more recent, coarse-grained census [DANE] has shown how those projections were outdated and under-estimated. The new numbers only indicate a 30% growth of the population of the municipality of Cali, including the inner center and the peripheral districts, which was uniformly taken into account for the calculation of the references. However, it is very likely that the population size increased more in the peripheral districts (therefore in $Z_1$) than in the city center ($Z_0$). Unfortunately, no more granulated data is available

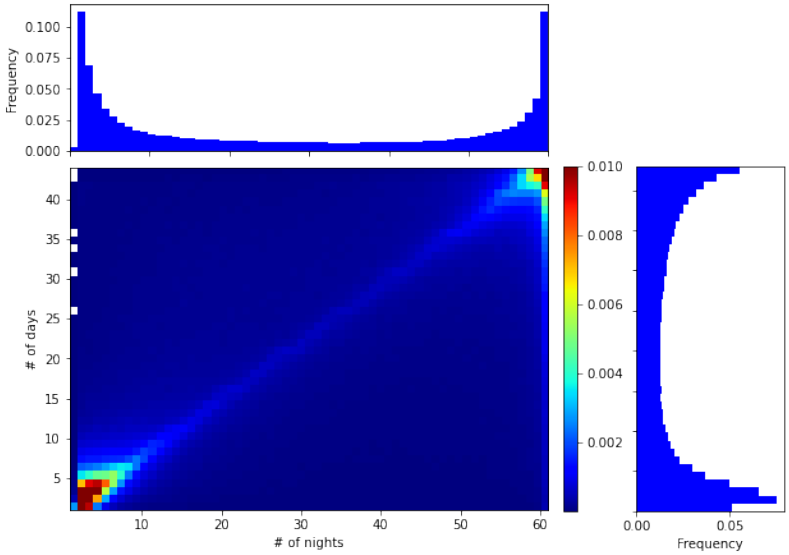
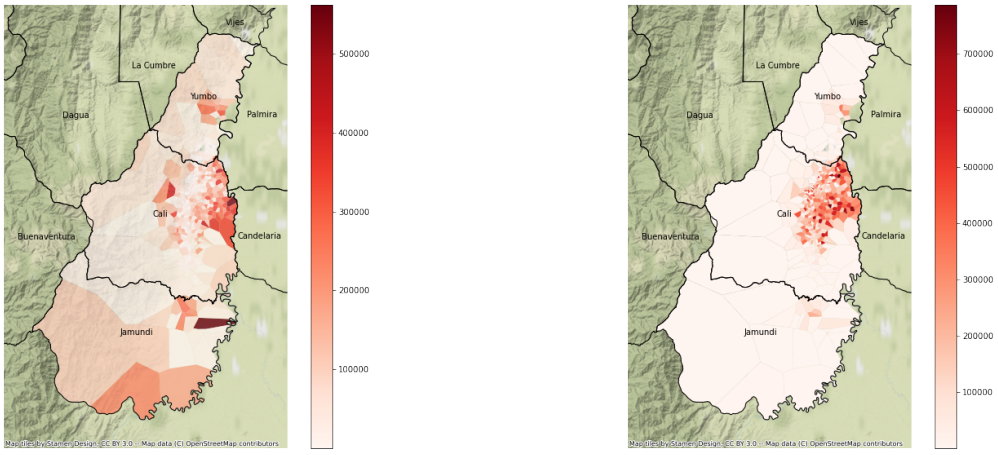
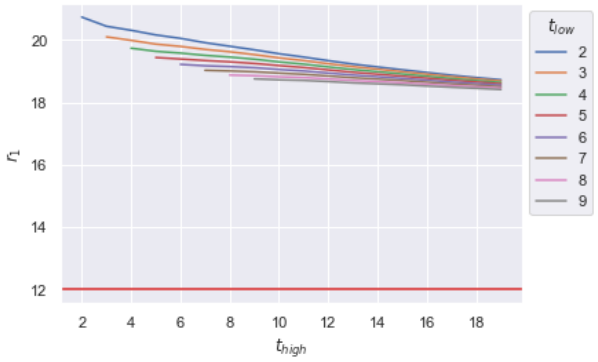



Figure 2.3: Heat map representing the amount of users observed for a given couple ($f_{night}$ in $Z_0 \cup Z_1$, $f_{day}$ in $Z_0 \cup Z_1$) for potential residents of $Z_0 \cup Z_1$.

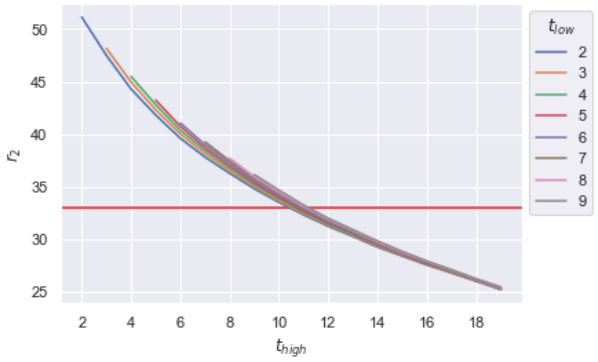



(a) (b)

Figure 2.4: Distribution of potential residents over the area. (a) Number of users per cell. (b) User density per cell (# of potential residents per km$^2$)




(a)

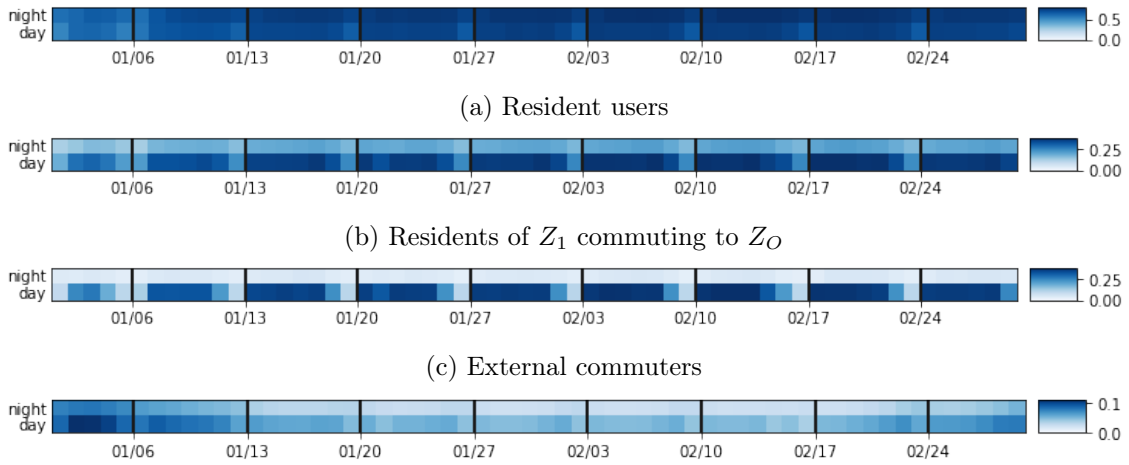



(b)

Figure 2.5: Exploration of parameters $t_{low}$ and $t_{high}$: (a) in relation with $r_1$. (b) in relation with $r_2$

| Ratio | $\frac{\lvert R_1 \rvert}{\lvert R_0 \rvert}$ | $\frac{\lvert C_1 \rvert}{\lvert R_1 \rvert}$ |
|---|---|---|
| Reference value | 0.12 | 0.33 |
| Observed value | 0.19 | 0.33 |

Table 2.5: Macroscopic indicators for parameter set (7, 10)




(a) Resident users

(b) Residents of $Z_1$ commuting to $Z_O$

(c) External commuters

(d) Visitors

Figure 2.6: Average individual call profile over the historical period for the different subcategories.

yet to characterize such local differences in the population growth compared to the latest fine-grained projections to confirm this justification. Based on Figure 2.5b, we set $t_{high}$ to 11. Then, we set the parameter $t_{low}$ to 7, considering than users being observed more than one week at night in the area must be considered as resident users rather than commuters (*cf.* Figure 2.1). This results in the ratios reported in Table 2.5.

Validating a user categorization method is difficult, as communication providers do not share the users' socioeconomic data for privacy reasons. However, we can explore the statistics of the different groups or their mobility dynamics to confirm the correct classification of users at an aggregated level. In Figure 2.6, we display the average individual call profiles of the users of each derived category per day and time window. On the x-axis are represented days of a two-months period. Each vertical black line represent the start of a new week (on Mondays). On the y-axis are represented the night and day time windows that divide a 24-hours day, the day cell aggregating the three early morning, restricted day, and late afternoon time windows. The color gradient of each cell represents the share of the users category, observed in $Z_0$ during the considered time window. Those average personal profiles on the historical period meet our expectations of the typical users of each class. Figure 2.6a represent the average call profile of residents of $Z_0$. Users are mainly observed during night and day, while we observe lower activity levels during the early morning and late afternoon transition time windows. We also observe a weekly seasonality with lower activities during weekends. Figure 2.6b presents the average historical profile of commuters residing in $Z_1$ while Figure 2.6c displays the profile of the external commuters. Both historical profiles are very similar. However, we observe that local commuters have a more diffuse presence during nights, early morning, and late afternoons. We can assume that the smaller distances users need to travel to reach the city of Cali explains those more diffuse behaviors. Those historical profiles also display weekend activity drops. Lastly, Figure 2.6d presents the average historical profile of the category of visitors. We can make several observations about this profile. First, the profile suggests more visitors

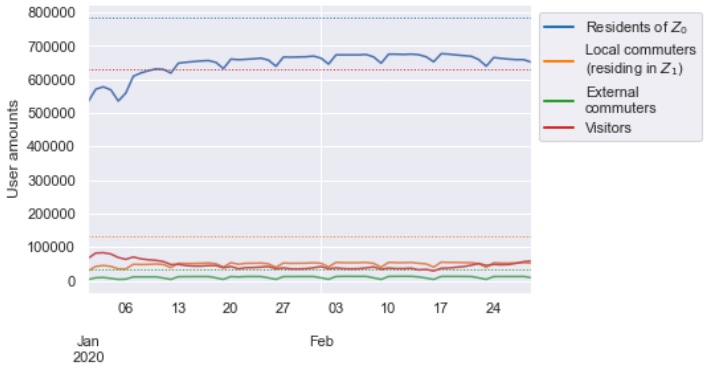



Figure 2.7: Daily presence trends of categorized users over the historical period

at the beginning and end of the historical period. The higher visitor activity during the two first weeks of the historical period is very likely related to the new year holidays and celebrations. During this period, the visitors seem to adopt a residing profile, with activity both during day and night. However, the increased activity at the end of the historical period is likely the consequence of a simple edge effect. The interruption of the history at the end of February makes the stays that have just begun seem shorter than they are. This interruption translates into an increase in the number of visitors detected during this period compared to the rest studied history. In the center of the historical period, we observe that visitors tend to have a daily commuting pattern rather than a resident pattern. We also observe a different weekly pattern compared to the rest of the population, with an increase of the visits over the week until Saturday when the visitors are more numerous.

Figure 2.7 summarizes well the daily trends of those different categories. We mainly observe the weekly seasonality of the different user categories and the behavior change at the beginning of the year. In this figure, we also display in dotted lines the cardinality of each group over the whole history. We identify 784,628 residents of $Z_0$, 131,712 residents of $Z_1$, 32,947 external commuters and a total of 628,470 visitors over two months. The daily amount of observed residents and commuters is close to the total historical group cardinality. However, the cardinality of all visitors is much higher than the daily number of visitors observed. This observation reflects a high turnover among this population, with many users observed for brief periods. Yet, the total number of visitors is high. This category very likely includes local users with scarce mobile phone use, making them indistinguishable from real visitors. The online integration of new days of data will allow to reclassify some visitors into resident or commuting users. It also should be noted than the application from time to time of a re-calibration of the method on an increasingly large dataset will allow to iteratively improve the results.

The various differences between total population and daily observed population also illustrates the daily dynamics of the populations and echoes the issue of temporal horizon of the analysis we raised in Section 2.1.2.4. In Figure 2.8, we display the increase of the observed population over the historical period, relatively to the total population size, for residents, commuters and visitors. 95% percents of the resident users are observed after 29 days. The same ratio is reached for the commuters population after 39 days, and 54 days for the visitor users. This seems to indicate that one month of data is a very minimum for categorizing users and especially to relate on the data observed for up-scaling purposes.

In the next paragraph, we further discuss the total population numbers and compare them to available baselines.

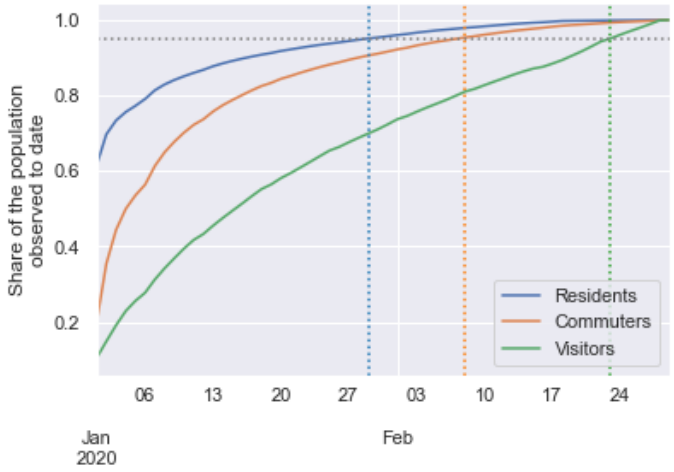



Figure 2.8: Daily presence trends of categorized users over the historical period

#### 2.1.3.2 Scaling factor estimation

The challenge when comparing the sampled population with the baseline data is to ensure the consistency of the two data sources. Ideally, this consistency should be both temporal and spatial. However, the low frequency of census and surveys limits the temporal comparability of the two data sources. The specific spatial scales of surveys and census are also a limit to spatial comparability. In our specific case, the origin place of users such as visitors and commuters is also unknown. We will address this problem by making assumptions about these geographical origins.

For residents, as no recent and fine-grained census data is available for the city of Cali, we ensure the temporal consistency of the ground truth with the mobile phone data by using population projections for 2020 [Departamento Administrativo de Planeación Municipal, 2018, Armitage Cadavid et al., 2019] based on data from the 2005 national census. The spatial scale of those projections is quite aggregated: it corresponds to *comunas* (districts) in large cities and *corregimientos* (rural townships) otherwise. Thus, we distribute the observed population of a BS cell in the different census blocks it intersects. This distribution takes into account population and area criteria according to the method developed in Bachir [2019]. Comparing this distributed sample population with the 2020 population projections, it becomes possible to measure the sampling ratios between sampled users and census population at the census block scale. This distribution results in population distribution displayed in Figure 2.9. Appendices A.1 also includes Table A.1 which details all scaling factors numerical values.

These values vary between 0.53 in El Hormiguero to 9.49 in Villacarmelo. The scaling factor is below 1 in El Hormiguero and La Elvira, and between 1 and 2 in the Comuna 22 of Cali, La Castilla, El Saladito, La Buitrera and Pance, which are administrative zones located in the north west of Cali, and in the South of Cali. Such low scaling factors either mean that some of the individuals' home locations assigned to these areas are wrong or that the census populations of these areas are underestimated. The first hypothesis may especially explain the highest sampling ratios in the northwest of Cali (La Elvira, La Castilla, and El Saladito). Indeed, the base stations there are higher than in the city and could capture calls from relatively long distances, resulting in residents with wrong house locations. However, this cannot explain the high sampling ratios observed in the city's south, characterized by a relatively low altitude. There, the hypothesis of underestimation population projections is more credible. It seems especially confirmed by the recent publishing of the conclusions of the 2018 national census carries at the *municipio*

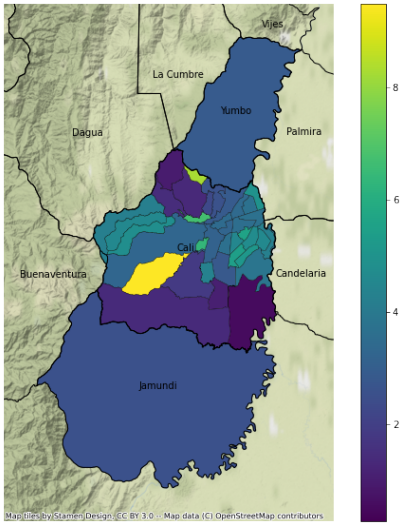



Figure 2.9: Scaling factors at the census block scale

scale and the subsequent update of the population projections at the same scale [DANE], that we already mentioned in the previous paragraph. At this scale (which covers the *comunas* and *corregimientos* together), the new population projections for 2020 in Cali's municipality show a 30% [DANE] growth of the global population compared to the previous projections, suggesting that the city grew in the last years faster than expected. To take this growth into account, we applied a uniform factor 1.3 to the population references. However, we believe that it might be concentrated in Cali's suburbs, and especially in the south of the city (Communa 22 and El Hormiguero). Indeed, the city could have significantly grown southward, where a lot of universities have settled. We expect the next finer-grained population estimates generated by the city of Cali to confirm or deny this hypothesis, and allow finer scaling factor estimations.

For commuters, we are making the assumption that they all live in Cali's metropolitan area, which includes, beside Jamundi and Yumbo, Candelaria and Palmira. The 2015 mobility survey [Metro Cali, 2015] reports 50,806 daily trips from Candelaria, Palmira and the nearby international airport Alfonso Bonilla Aragon. Considering the distances between those municipalities and the city of Cali, we can safely assume that it is close to corresponding to 50,806 daily travelers. The categorization of users resulted in a total of 32,947 commuters identified, which results in the scaling factor below:

$$f_C = \frac{50{,}806}{32{,}947} = 1.42 \tag{2.11}$$

This scaling factor is relatively low. This can be explained by several factors. First, our categorization may result in misclassified residents, that are interpreted to be commuters because of a weaker communication activity at night. Second, the reference data is provided for trips coming from the municipalities centers and do not take into account the rural areas around that may contribute to the commuter population. Third, we are assuming here that the commuter population is strictly coming from the metropolitan area of Cali, and from its urban centers, while the employment area of Cali could spread further around these zones.

When it comes to visitors, the reports generated by the tourism service of the city of Cali and by the Colombian immigration service [Infometrika, 2019] show, for the year 2018, a total number of foreign visitors:

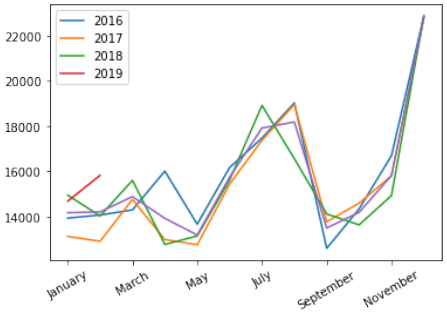



Figure 2.10: Scaling factors at the census block scale

$$|V_{year}^{ref,international}| = 184{,}473 \tag{2.12}$$

As already stated in 2.1.2.4, this number refers only to travelers whose principal destination is Cali. Consequently, this number does not include travelers visiting Cali but spending more time in another city. In 2018, according to the same source, foreign visitors represented 13.3% of the visitors of the city, the other 86.7% corresponding to national people visiting yearly.

$$|V_{year}^{ref,national}| = 1{,}202{,}542 \tag{2.13}$$

The 2018 touristic report [Secretaría de Turismo de Cali, 2019] also provides information on the foreign visitors flows throughout the year. These trends are displayed in Figure 2.10.

In 2018, we estimate that the foreigner visitor population over the historical period (January and February) covered 15.48% of the yearly visitors. Assuming that national and international visitors follow the same influx trends, and applying Equation 2.8 on this period, we can estimate the national visitors population size over the same period:

$$|V_{hist}^{ref,national}| = 0.1548 \cdot |V_{year}^{ref,national}| = 186{,}154 \tag{2.14}$$

This number is far below the visitor population observed from our categorization of the data (|V| = 628,470). In Equation 2.9, this results in a scaling factor smaller than one:

$$f_V^{hist} = \frac{186{,}154}{628{,}470} = 0.27 \tag{2.15}$$

We explained in Section 2.1.3.1 that this category is likely to include more users than it should, considering that many users that are barely active will be considered as visitors, and that it includes a larger range of visitors profiles than the one considered in the baseline number. Despite these limitations, this is the ratio we will apply to identified visitor users so as to represent the mobility of touristic national visitors in our mobility and traffic variables estimates.

### 2.1.4 Discussion

In this section, we have discussed a simple categorizing method to classify users according to three main mobility profiles (residents, commuters, and visitors.), and formalize up-scaling methods for the two categories usually neglected by the literature, the commuters and the visitors.

The categorization method we propose classify users based on binning rules relying on a few presence features. The binning parameters are calibrated thanks to macroscopic ratios between population subgroups derived from the ground truth population characteristics. This approach has the advantage of guaranteeing to a certain extent the reliability of the results, in a context where individual ground truth data are inaccessible. However, the observation of average communication profiles and population trends have returned results consistent with our expectations. Although easy to implement and to calibrate, this binning approach presents the disadvantage to categorize users into very strict and geometric categories, while the population is characterized by an important diversity and continuity of the presence features. In future works, the implementation of more sophisticated stepwise classification methods Furletti et al. [2012], Furletti et al. [2013], Mamei and Colonna [2018] may be an approach to handle the data continuity challenge. These methods rely on the supervised labeling of characteristic users (or archetypes) to classify the rest of the population. In our approach, the use of resident users with commuting behaviors was a key element of the calibration step. We look forward to apply learning algorithms on this population to further characterize the commuting behaviors and support a more flexible categorization method. Increasing the flexibility of the method can also be achieved by increasing the number of objective categories. However, this raises the question of getting reference population sizes for sample up-scaling and analysis expansion remains open.

In a step towards this direction, we formalized a method for up-scaling users according to the three objective mobility profiles. Such a formalization, and a standardization of the reference data, is lacking proper literature and results in mobility studies focusing on the easiest users to expand: residents of the monitored area. Although the available baseline data was limited, we make assumptions about the users profiles and geographic origins which support the estimation of categorical scaling factors. These assumptions are restrictive, and have as consequence that the mobility and traffic variables estimated *in fine* can only be representative of the population characterized by the reference data, such as touristic visitors and near-by commuters. In future works, refining the user categories, for instance to distinguish different categories of visitors, and focus on those for which baseline data is available, could be a way to increase the consistency of the population compared through the scaling factor estimation.

In this section, we have also emphasized the temporal dimension of categorization and scaling processes, which is often neglected in favor of spatial consistency. We especially gave an illustration of the sensibility of the users detection to the sample duration. Generally speaking, the largest the sample temporal depth, the better the user categorization. In practice however, such analyses are often limited by reduced temporal depths, or re-indexing of individuals. On the other hand, this observation also indicates that in the context of a continuous data reception in line with this work, regular re-categorization can allow to refine the users classes. This would positively impact the results in two ways, first through this improved user categorization, second through refined scaling factors that would better account for the internal dynamics of the population. However, regardless of the length of time studied, and the progressive refinements of the classification, its results cannot be accurate without quality census and survey data, which are the cornerstone of sample adjustment.

## 2.2 Study of the road network

### 2.2.1 Introduction

We have previously outlined how the specific characteristics of the CDRs data affect the reconstruction of trajectories and thus travel distances and regional flows.

First, due to the data temporal sparsity, it is very unlikely to observe the complete trajectory of a user between their origin and destination. If only segments of that trajectory are observed, the users appears to be missing in the region they traveled while being inactive, and the flows in those regions are underestimated. Therefore, there is a need for a map-matching process, to infer the region the user traveled when no data was collected. Several literature works have explored how to complete partial CDR trips, comparing user footprints with their historical data or with other users' Asgari et al. [2016], Bonnetain et al. [2019]. However, these methods can be time-consuming when processing large sample size, and are sensitive to the communication rates of the considered users.

Besides, the spatial imprecision and low resolution of the data raise the challenge of the estimation of traveled distances. The raw data only enables the calculation of cell-to-cell distances. Several studies, related to the estimations of traffic-related emissions, have relied on this simple approach Li et al. [2016]. However, it remains highly approximate, and does not render the road network impact on the trip lengths nor is adapted to the calculation of traveled distances within cells.

The reconstruction of the trajectories of individuals, and the estimation of the distances associated with these trajectories is therefore a major issue of this thesis. Considering that CDR data are spatially insufficient to answer this question alone, we resort to answer it to a literature about the Macroscopic Fundamental Diagram (MFD) theory Daganzo [2007], Geroliminis and Daganzo [2008]. This framework includes several studies aiming to characterize trip distance at a sub-regional level, *i.e.*, based on paths defined as crossing several successive urban regions. In particular, some works [Batista et al., 2021b] are about automatic identification of main (*prevailing*) paths in the system based on the road network analysis. It has been validated using GPS data in Lyon. During my PhD, I have contributed to this study. However, such a method does not directly provide trip length estimates. This is why in this chapter, we resort to another approach [Batista et al., 2019] based on artificial trips sampling. This method relies on sampling many origin-destination pairs and systematically browsing the network with a shortest path algorithm. In this section, we propose to adapt it to the large network of our case study.

Section 2.2.2 details the original algorithm, and propose two major improvements to make it scalable to larger networks. Section 2.2.3 presents the results that the method provides on our case study.

### 2.2.2 Methodology

#### 2.2.2.1 Preliminary definition

**Definition 1 (Regional path)** *Considering a regional partitioning $\mathscr{R}$, a regional path $p$ [Yildirimoglu and Geroliminis, 2014, Batista et al., 2019] is defined as a sequence of adjacent regions $p = (r_i)$, $r_i \in \mathscr{R}$. We further use $\mathscr{P}$ to refer to the set of possible regional paths, and $\mathscr{P}_r$ to refer to the regional path that cross a region $r$:*

$$\mathscr{P}_r = \{p \in \mathscr{P} | r \in p\} \tag{2.16}$$

#### 2.2.2.2 Problem statement

Considering that CDR data are inadequate for an individual analysis of travel distances, we propose to estimate regional traffic volumes on the basis of a decoupling of flow and average regional distances:

$$TTD_r(t) = q_r(t) \cdot \bar{L_r}(t) \tag{2.17}$$

where $q_r$ is the regional flow in $r$ and $\bar{L_r}$ the average distances in $r$, that we will further assumed to be static. Because regional average distances have been shown to vary with path [Batista et al., 2019], we incorporate this distinction into our analysis for greater accuracy in the estimation of the traffic volume:

$$TTD_r(t) = \sum_{p \in P_r} TTD_{p,r}(t) = \sum_{p \in P_r} q_p(t) \cdot \bar{L_r}, p \tag{2.18}$$

where $P_r$ is the set of paths crossing $r$, $q_p(t)$ is the path flow along $p$ at $t$, and $\bar{L_{r,p}}$ the average distance. Therefore, the problem underlying the estimation of traffic volume is of two kinds:

1. How can we reconstruct the individual trajectories in order to evaluate accurate path flows $q_p(t)$ ?
2. Howe can we estimate estimate the region- and path-specific average travel distances ?

The work presented in the following section provides partial answers to this question.

#### 2.2.2.3 Presentation of the reference work

In Batista et al. [2019], the authors provide a method that relies on a simple and systematic network analysis for identifying the main regional paths in a multi-regional network, along with estimating the average regional travel distances. It relies on the assumption that the shortest paths represents well the distance traveled by users, hypothesis that will also be ours in this thesis.

Let us consider a road network as a directed graph $G = (V, E)$. The method is first based on the uniform sampling of a set of network nodes, coupled into a set of origin-destination couples $N_{od}$. For each couple $od$, the authors proceed to the computation of the shortest path $sp(od)$ from the origin to the destination, using Dijkstra's algorithm. This results in a set $T$ of *virtual trips*. Although the authors consider different resolutions for averaging distances, we focus the regional path level, which consists in estimating $\bar{L_{r,p}}$ the average travel distance in region $r$ along path $p$:

$$L^{-}_{r,p} = \frac{\sum_{k \in T} \delta_{k,p} l_{k,r}}{\sum_{k \in T} \delta_{k,p}} \tag{2.19}$$

where $\delta_{k,p}$ equals 1 if the virtual trip $k$ travels along the regional path $p$, 0 otherwise, and $l_{k,r}$ is the distance traveled by $k$ in region $r$ along $p$.

Besides, for each regional origin-destination couple, the method application allows identifying a set of prevailing regional paths $\mathscr{P}_{OD} = \{p = (r_1, ..., r_n) \in \mathscr{P} |\ r_1 = O \wedge r_n = D\}$ that provide the shortest distances to travel from $O$ to $D$.

Applying this method on our case study presents a two-fold advantage. On the one hand, the exploitation of the regional paths can provide an insight into the paths traveled by users and support a regional map matching of individual trips. On the other hand the regional average trip lengths can be used as a reference for the total traveled distance estimation. Thus, this approach seems suitable to fill the gaps in the cell phone data for the purpose of estimating the traffic volume.

However, so far, the method has only been applied to small networks (757 links and 431 nodes, Batista et al. [2019]), because it is time consuming. As our case study is very large (28,558 nodes), applying the method of the literature would be too costly. Therefore, we propose two central adaptations of the method to fasten it. These adaptations are presented in the following section. For further simplicity, we will refer to the literature method as $M1$, and to our enhanced method as $M2$.

#### 2.2.2.4 Algorithm modifications

The first adaptation of the method relies on the property of the Dijkstra shortest path algorithm, selected by the authors, which actually returns the list of shortest paths to all the nodes of the graph. $M_1$ relied on the sampling of origin and destination nodes, and the iterative running of the shortest path algorithm for each *od*. Instead, we suggest sampling a set of vertices $N_v$. Based on this sampling, we define a new origin-destination set, as:

$$N'_{od} = \{(o, d) \in N_v \times N_v \setminus \{o\}\} \tag{2.20}$$

$N'_{od}$ corresponds to the set of all possible origin-destination couples derived from $N_v$. Iteratively, each node is labeled as origin. Let $N'_{v_0} = \{(o, d) \in N'_{od} |\ o = v_0\}$ be the subset of origin-destination couples from $v_o$. Running Dijkstra's algorithm once from $v_0$ returns the shortest paths from $v_0$ to every other nodes of the networks. Therefore, it results in $T_{v_0}$ the set of virtual trips for every origin-destination couple in $N'_{v_0}$. As a consequence, $|N_v|$ runs of Dijkstra's algorithm allow to compute a set of virtual trips of size $|N_v| \times (|N_v| - 1)$.

Figure 2.11 illustrates the differences between these two approaches, by displaying the virtual paths derived at each shortest path calculation. For the sake of readability, only two iterations where displayed for the method $M_2$. As shown in Table 2.6, for an equal number of Dijkstra's operations ($n$), applying method $M_2$ allows cutting by 2 the number of sampled points compared to $M_1$, while multiplying by a factor $n - 1$ the number of considered *od* pairs.

The second adaptation of the method focuses on the node sampling. In Batista et al. [2019], the authors study the sensibility of the method to the number of the *od* sampled. The larger the sampling rated, the larger the network coverage and the better the results. After a sensibility analysis, the authors recommend a sampling 10,000 origin-destination pairs, for a network of 757 links and 431 nodes, in order to ensure a good coverage of the network links and accurate estimation of the lengths. Our network being much larger

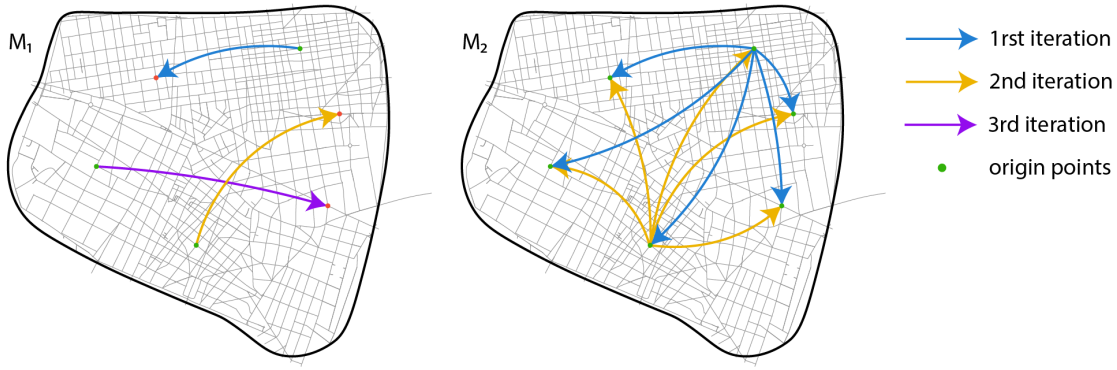



Figure 2.11: Computed shortest paths per iteration in $M_1$ and $M_2$

| | M1 | M2 |
|---|---|---|
| Sampled points | 2n | n |
| Considered *od* pairs | n | n(n-1) |
| Dijkstra runs | n | n |

Table 2.6: Comparison of sampling and running numbers

(28,558 nodes and 72,648 links), we propose to drastically reduce the candidate nodes for sampling, by focusing on the most strategical ones. We make the following observation. For a given regional origin-destination couple, there are many possible origin-destination couples at the microscopic level. However, the routes between all these *od* have in common that they all cross the borders of the region $O$ and $D$. Therefore, we propose to simplify the regional length calculation proceeding with three steps:

1. We start by focusing on estimating border-to-border regional distances.

2. Separately, we estimate inner regional distances, based on both in-going and out-going directions.

3. Lastly, we chain the border-to-border distances with the inner distances within the origin and destination regions.

Let $b_{r_i,r_j}$ be the set of border vertices between two regions $r_i$ and $r_j$ of $R$, and let $B$ be the set of vertices. We set $N_v = B$, and compute the shortest paths from each border node to all the others. This results in a set of virtual border-to-border virtual trips $T_b$. Let us consider a regional path $p = (r_1, ..., r_n)$. We further call a border-to-border path $p_{o,d}$ a regional path $p$ extended by the origin border $b_{r_o,r_1}$ and the destination border $b_{r_n,r_d}$: $p_{o,d} = (b_{r_o,r_1}, r_1, ..., r_n, b_{r_n,r_d})$

We define the border-to-border regional length estimate in $r$ along the path $p_{i,j}$ as:

$$\bar{L}_{r,p_{r_o,r_d}} = \frac{\sum_{k \in T_b} \delta_{k,p_{o,d}} l_{k,r}}{\sum_{k \in T_b} \delta_{k,p_{o,d}}} \qquad \forall r \in p_{r_o,r_d} \tag{2.21}$$

It is important to note that Equation 2.21 does not define a regional length estimate in $r_o$ and $r_d$. This is the objective of the second step of the method.

In this second step, we focus on the computation of the regional inner distances. Each region $r$ of $R$ is considered in turn, and we apply, for each, the following process. Firstly, let $b_r$ bet the set of border nodes of $r$, and let $G_r = (V_r, E_r)$ so that $V_r \in r$ and $E_r \in r$

be the subgraph of $G$ over region $r$. We proceed to the uniform sampling of a set of nodes $N_v^r$ from $r$.Then, we define two origin-destination sets, as follows:

$$N_{out}^r = \{(o,d) \in N_v^r \times b_r\} \tag{2.22}$$
$$N_{in}^r = \{(o,d) \in b_r \times N_v^r\} \tag{2.23}$$

Running Dijkstra's algorithm in $G_r$ on those set results in two sets of virtual inner trips $T_{out}^r$ and $T_{in}^r$. Further considering $A_r$ the set of regions that are adjacent to $r$, we estimate the outgoing distance to a border $b_{r,r_i}$, $r_i \in A_r$:

$$\bar{L}_{r,r_i}^{out} = \frac{\sum_{k \in T_{out}^r} \delta_{k,b_{r,r_i}} l_k}{\sum_{k \in T_{outgoing}^r} \delta_{k,b_{r,r_i}}} \tag{2.24}$$

and the incoming distance from the border $b_{r,r_i}$:

$$\bar{L}_{r,r_i}^{in} = \frac{\sum_{k \in T_{in}^r} \delta_{k,b_{r_i,r}} l_k}{\sum_{k \in T_{outgoing}^r} \delta_{k,b_{r_i,r}}} \tag{2.25}$$

This step-wise, direction-differentiated approach is necessary as the distance traveled within the region may vary if the links are bi-directional or not.

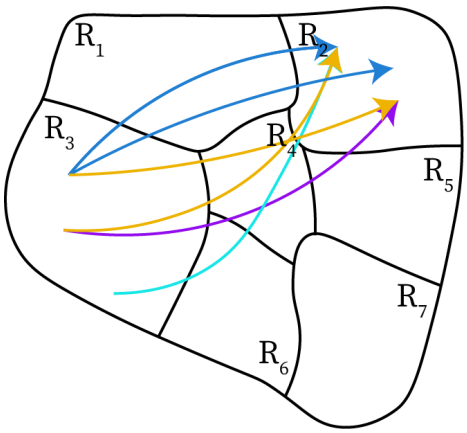
Synthetizing the Equations 2.21, 2.24 and 2.25, we define, for a given full path $p = (r_1, ..., r_n)$, the regional distances along $p$ as:

$$\bar{L}_{r,p'} = \begin{cases} \bar{L}_{r_1,r_2}^{out} & \text{if } r = r_1 \\ \bar{L}_{r,p_{r_1,r_n}} & \forall r \in p \setminus \{r_1, r_n\} \\ \bar{L}_{r_{n-1},r_n}^{in} & \text{if } r = r_n \end{cases} \tag{2.26}$$

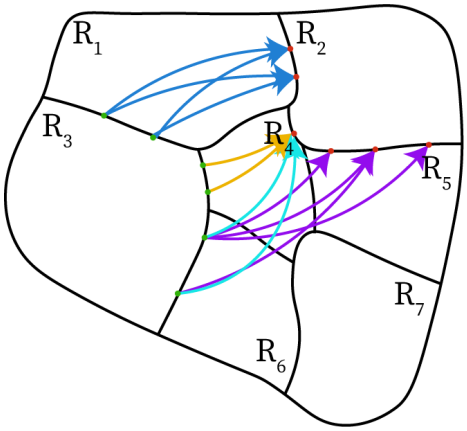
$$\bar{L}_{r,p'} = \begin{cases} \frac{\sum_{k \in T_{out}^r} \delta_{k,b_{r_1,r_2}} l_k}{\sum_{k \in T_{outgoing}^r} \delta_{k,b_{r_1,r_2}}} & \text{if } r = r_1 \\ \frac{\sum_{k \in T^b} \delta_{k,p_{r_1,r_n}} l_{k,r}}{\sum_{k \in T^b} \delta_{k,p_{r_1,r_n}}} & \forall r \in p \setminus \{r_1, r_n\} \\ \frac{\sum_{k \in T_{in}^r} \delta_{k,b_{r_{n-1},r_n}} l_k}{\sum_{k \in T_{outgoing}^r} \delta_{k,b_{r_{n-1},r_n}}} & \text{if} r = r_n \end{cases} \tag{2.27}$$

Figure 2.12 illustrates this process. The advantage of this second modification of the algorithm is two-fold. On one side computing border-to-border shortest paths allows to significantly reduce the number of nodes considered and the number of Dijkstra's runs, while focusing the computation on necessary nodes from the origin to the destination region (Figure 2.12b).On the other side, the regional restriction of the network to $G_r$ allows computing with increased accuracy inner incoming and outgoing average distances from a large node samples without while containing the computation time (Figure 2.12d).

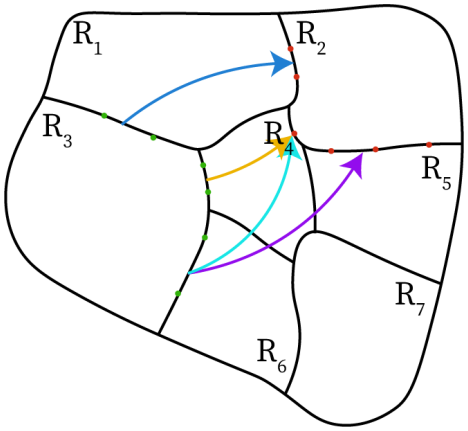
However, it should be noted that the results provided by our approach are likely to differ from the original method. In fact, the step-wise adaptation of the method may result in some biases in the estimation of trip lengths within regions. In $M_1$, the distances in border-to-border regions are likely to be sensitive to the relative importance of origin and destination border nodes: as a consequence each shortest-path between an origin border node and a destination border node has a specific apparition frequency. In our approach instead, each virtual trip characterizing border-to-border path is given the same weight in



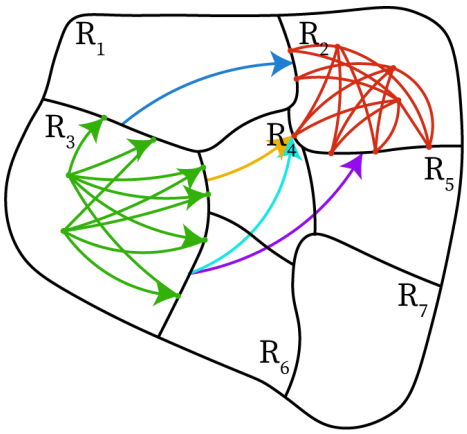

(a) Original method



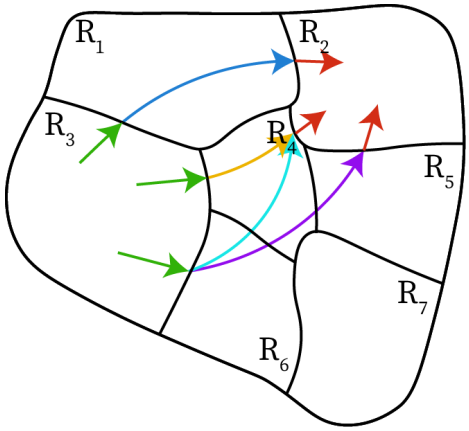

(b) Border-to-border shortest paths calculation




(c) Border-to-border average trip lengths calculation




(d) Inner shortest paths calculation




(e) Average regional paths trip lengths reconstruction

Figure 2.12: Schematic illustration of the second modification of the literature method.

the average distance calculation. Considering the limit width of the regions and regional paths, we will assume a low variability of the distances from one border node to the other. This bias may have a stronger impact on the calculation of inner distances within the origin and destination region. For instance, the outgoing distance in $r_i$ towards border $b_{r_i,r_j}$ is calculated as the average distance of all sampled inner nodes toward the border nodes. In the reality, a specif border might be more attractive to the closest inner nodes than the further ones. If so, our inner distances are likely to be over-estimated. Our objective being to provide a cost-efficient alternative to the literature approach, we have not investigated this subject further yet, but these analyses will be the subject of future specific studies. In future work, we may consider weighting, or focusing on border nodes based on their level of service or their location compared to the origin-destination azimuth for increasing the cost-efficiency of the method and the accuracy of its results. Another ambition would be to take into consideration the local demand, to generate demand-consistent average distances.

### 2.2.3 Results

This section illustrates the results provided by the method application to the city of Cali. Those results are presented for the regional network $\mathscr{R}_1$.

The computing of border-to-border shortest paths results in 5,616,900 different routes describing 218,273 different border-to-border regional paths.

Once the border-to-border regions have been prolongated with the origin and destination region inner lengths, we get 256,550 different regional paths. On average, we observe 11.67 different paths per OD couple.

We now dig into the spatial dispersion of the regional paths observed. We define an indicator $I_C$ of the spatial cohesion of regional paths for a given OD as follows. Let $O$ be a regional origin, and $D$ a regional destination. Let $P_{OD}$ be the set of paths observed between $O$ and $D$, $R_{OD}$ the set of regions observed in $P_{OD}$. For each region $r$ in $R_{OD}$, we define $f_r$ the region crossing frequency:

$$f_r = \sum_{p \in P_{OD}} \delta_{r,P}, \qquad \delta_{r,P} = 1 \text{ if } r \in P,\ 0 \text{ otherwise.} \tag{2.28}$$

From this, we define $I_C(OD)$ as the average regional crossing frequency over $R_{OD}$, normalized by the cardinality of $P_{OD}$:

$$I_C(OD) = \frac{\bar{f}^{R_{OD}}}{|P_{OD}|} \tag{2.29}$$

Figure 2.13 displays the distribution of $I_C$ over the whole OD set. The mean value of the distribution is 0.66, which ranges between 0.27 and 1, this maximum value being reach for OD with a single prevailing path detected. In Figure 2.14, we illustrate the spatial dispersion of regional paths for five different OD couples. In Figure 2.15 displays instead OD couples with paths with a strong spatial cohesion.

Lastly, Figure 2.16 presents an illustration of the range of average trip lengths associated with different paths within a same region.

### 2.2.4 Conclusion

In this section, we propose making the most of a method of the MFD-literature [Batista et al., 2019] to construct a background knowledge on our case-study's road network. Given

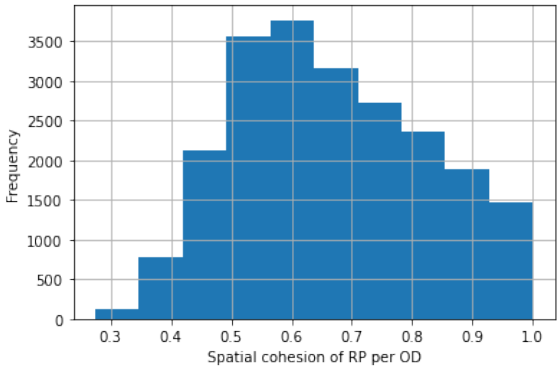



Figure 2.13: Distribution of the prevailing paths dispersion level per OD

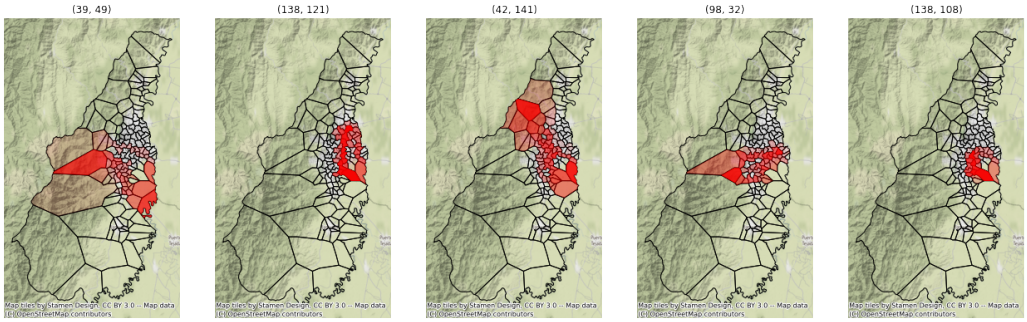



Figure 2.14: Distribution of the prevailing paths dispersion level per OD

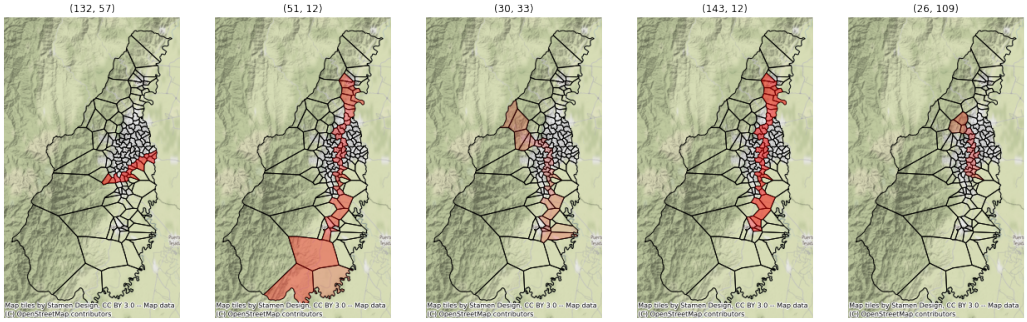



Figure 2.15: Distribution of the prevailing paths dispersion level per OD

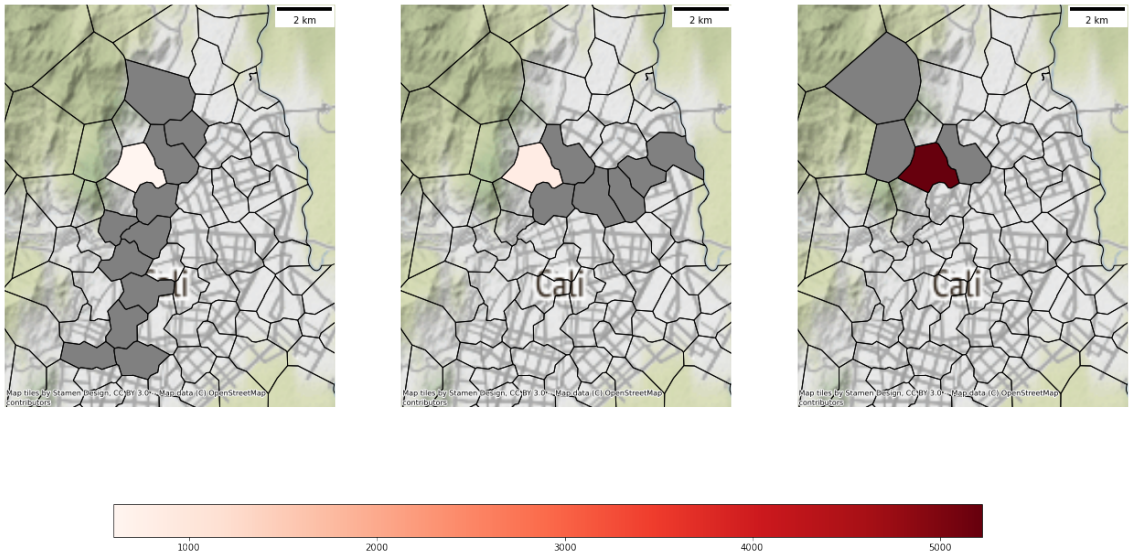



Figure 2.16: Distribution of the prevailing paths dispersion level per OD

a predefined regional partitioning of the road network, this method relies on the systematical analysis of the network and results in the identification of the prevailing regional paths between each OD and the estimation of the typical regional lengths along those paths. The application of the method as it stand in the literature can be time consuming with dealing with large networks. Therefore, we proposed two structural modifications of the chore algorithm. Those modifications allow us to run the method on our road-network. The application of this method to our case-study provide us with a valuable insight into the network specific characteristics. But most importantly, it returns two kind of results that are essential to our framework. First, we identified the prevailing paths between a regional origin and destination. These prevailing paths will later be the baseline to inferring the paths traveled by users, and therefore to estimate appropriate population flows. Second, the knowledge about the regional average traveled distances, once crossed with the flows, will provide an estimation of the total travel distances.

## 2.3 Conclusion and perspectives

This chapter gathers two essential analyses for applying our framework to an urban area and meet our ambition to estimate traffic volumes as accurately as possible. Reaching this ambition requires taking up various challenges.

First, it requires considering the overall urban population and its large behavior ranges. Many literature studies have focused on resident users for sake of simplicity. However, the conclusion drawn from these studies can only be representative of this share of the urban population, but in no way can account for the entire urban population, which includes other types of urban users. In this work, we aim to be more exhaustive and consider other mobility actors, such as external commuters and visitors. The first section of this chapter goes in this direction, with an in-depth analysis of the population observed in the area. We perform a categorization of users according to three typical mobility profiles as well as a spatialization of resident users with the estimation of their home location. On these bases, we estimate scaling factors for the various population categories. Theses scaling factors will allow to expand our sample and upscale the related traffic variables.

Second, the scarcity of CDR data raises the question of the complete identification of flows, and the correct estimation of related average distances. So far, the CDR-based

literature has adopted either too simplistic or too costly methods for answering these problems. Therefore, we propose the adaptation of method developed in the context of the MFD literature to answer both. We make two essential modifications of the original algorithm, and suggest another, to make that method less time-consuming and easily scalable. Then, applying it to our case study, we draw essential conclusions concerning the prevailing regional paths as well as the average trip lengths along these regional paths. These conclusions will provide insight into the CDR-derived mobility observation, and will form the basis for the reconstruction of the traffic variables of interest.

# Conclusion of Part I

In this part of the thesis, we have laid the groundwork for mobility analysis in our case study, the city of Santiago de Cali, Colombia.

In Chapter 1, after presenting this study area, we introduce the available cell phone data and their characteristics. These CDR data have the particularity of being dependent on the communication activity of the users. Moreover, their spatial granularity depends on the communication network density and thus translates into a variable resolution over the area. To ensure a more homogeneous level of analysis and guarantee a certain level of representativeness of the local subsamples, we proceed to the aggregation of the Voronoi cells, according to two criteria, population and surface. This step results in a regional partitioning used as a reference in the rest of our work.

In Chapter 2, based on the spatial partitioning previously developed, we conduct two preliminary but necessary studies to reconstruct mobility from our data. The first analysis concerns the sampled population and aims at a global estimation of the flows. While most literature focuses on a resident population to derive mobility indicators, we highlight the limitations of this approach when evaluating global indicators intended for evaluating atmospheric emissions. In response to this issue, we propose categorizing the population according to their presence profiles (residents, commuters, visitors) and formalizing methods for upscaling these sub-samples. The second analysis is based on the selected regional division and a proposal for automatic road network routing to determine, on the one hand, the typical regional paths between an origin and a destination, and on the other hand, the average regional lengths according to these different regional paths. This preliminary analysis of the network will support the reconstruction of the trajectories taken by the users and the estimation of the total distances traveled.

The main contributions of this part are the following.

- We propose basing global mobility on a population classification for the first time. This classification will support an exhaustive mobility analysis by considering broader sections of the population than the literature.
- We formalize subsample expansion methods that relate sample categorization and global population representation.
- While the CDR data lacks the granularity for a complete and accurate traveled path and distance estimation, we adapt a literature method to identify the prevailing regional paths and estimate average traveled distances at a large urban scale.

In future works, we could consider improving some of the analyses conducted in this section by:

- Refining the spatial division of the network. Besides ensuring minimal areas and sample sizes, one could consider clustering base stations with similar communication dynamics or consistent underlying road networks.
- Improving the categorization of the population by considering new user classes and applying learning methods.
- Extending the systematic analysis of the road network to other modes of transport.

# Part II

# Mobility patterns reconstruction

## Introduction

This part of the thesis focuses on the estimation of traffic volumes from CDR data. It proposes a distinct mobility reconstruction approach for two main groups of individuals. These groups are established partly on the basis of the categorization performed in the chapter 2, partly on the basis of the regularity of the individuals' mobility patterns.

The first categorization classified users according to their observed patterns of presence in the city. This results in three groups of individuals: residents, commuters and visitors. These three categories can be aggregated into two main categories, local users and non-local users. A key difference between these individuals is the depth of data history. While visitors have a high turnover rate and relatively short identified stays, residents and commuters have much more extensive data histories. The availability of these histories, which is enhanced by the fact that user identifiers do not change, provides a considerable resource for interpreting and enriching mobility.

Indeed, human mobility is controlled by a number of habits that make it very regular, both on an individual and collective scale. When using parcel data such as cell phone data, this regularity can be a very important asset for interpreting and enriching mobility. For example, if the data for a day is fragmented, historical information can be used to supplement it. This is valid provided that the individual is indeed mobile on a regular basis. If this characteristic is common to a large part of the population, there are outliers who, because of their professional activity for example, will show erratic mobility.

In this part of the thesis, we propose to distinguish these individuals according to the regularity of their mobility patterns, and develop two mobility reconstruction methods adapted to each category. This distinction between regular and irregular individuals is made according to two principles. First, all visitors are considered as irregular individuals. Second, the separation of premises into regular and non-regular premises is based on the measure of their entropy.

The entropy of a user can be measured in different ways and take more or less into account the context of mobility. The reader can refer to the work of Song et al. [2010] on this subject. In order to classify local individuals into regular or irregular individuals, we use the *temporal-uncorrelated* entropy:

$$S_u = -\sum_{i=1}^{N_u} p_u(i) \cdot log_2 p_u(i) \tag{2.30}$$

where $p_u(i)$ is the visit probability of $i$ in user $u$'s historical data. Based on this quantity, we define an arbitrary threshold beyond which an individual is considered irregular. In this case, we set this threshold at one standard deviation above the mean of the entropy. All individuals whose entropy is higher than this threshold value are considered as irregular users, those whose entropy is lower are considered as regular users. Depending on whether the individual is a regular local, or an irregular local or a non-local (visitor), they will not be integrated in the same mobility reconstruction approach.

Table 2.7 summarizes the double categorization that results from the profiling of individuals according to their frequency of use of the area, and, for local individuals, from the measurement of their entropy. It also shows how the mobility of each category will be processed:

- For regular users, we propose a individual mobility reconstruction, based on their historical data and their mobility regularity. This approach is developed in Chapter 3.

- In view of poor or irregular mobility histories, the mobility of non-local and irregular local users may be more difficult to enrich. Therefore, we propose for these individuals a more aggregated reconstruction of mobility, based on the direct estimation of the total travel distances. This method is developed in the Chapter 4.

| | | Low pattern regularity | High pattern regularity |
|---|---|---|---|
| Local users | Resident<br>Commuter | Collective and aggregated analysis | Individual and regional analysis |
| Non-local users | Visitors | Collective and aggregated analysis | |

Table 2.7: User categorization for mobility analyses purposes

In practice, the categorization of local individuals according to their entropy requires the extraction of mobility from the CDR data sequences, and in particular of static phases or activities The mobility extraction we implemented was based on the literature [Jiang et al., 2013, Toole et al., 2015]. Appendix B.1 describe this extraction process. Here, we introduce the basic related definitions and the main lines of the approach.

**Definition 2 (Stay)** *We define a stay as a sequence of consecutive communication events generated by a user in a restricted perimeter, during a minimum duration.*

**Definition 3 (Potential stay)** *We define a potential stay any communication event that are not labeled as stays because of short duration, but that collocate with known stays.*

We extend these literature-derived definitions with the notion of *quasi stay*, that complement the previous concepts. Its purpose is to specifically identify commuters entering or leaving the area and to consider their entry or exit points as a proxy for their origin or destination. The detection of these quasi stays is further discussed in Appendix B.

**Definition 4 (Quasi stay)** *A quasi stay is defined as a communication event sufficiently distant from the previous stay and followed by a period of inactivity of several hours, or conversely, a communication event sufficiently distant from the following stay and preceded by a period of inactivity of several hours.*

We further call static phase, or observed activity, any detected stay, potential stay or quasi stay.

**Definition 5 (Pass-by points)** *We define a pass-by point any communication event that is neither a stay, nor a potential or quasi stay.*

The mobility extraction process follows the following pattern:

1. Identifying stays;
2. Stabilizing stays, *i.e.*, associating to the same position stays that are distant in time but spatially close.
3. Identifying potential stays and quasi stays;
4. Labeling remaining communication events as pass-by-points.

Outline

# Chapter 3

# Trip matching and Path Flow estimation

## 3.1 Introduction

Path flow estimation is critical to estimate total distances traveled at the regional scale. Paipuri et al. [2020] have demonstrated how cell phone data can be processed to derive path flow distribution, among other variables. The authors argue that because these data are massive and user-centered, they are better adapted to path flow estimation than the more traditional sources like floating car data. However, their work calls on a specific type of mobile phone data, Location-Based Network Services (LBNS) data, *i.e.*, GPS or wifi-based data generated by smartphone apps in the background. The temporal and spatial resolutions of these data are much higher than for CDRs. Therefore, they offer rich trajectories with precise position data, allowing to reconstruct path-flows quite easily.

Due to the temporal sparsity and low spatial resolution of CDR data, estimating path flows from this source is difficult. To the best of our knowledge, no study has dealt with this question yet. So far, the literature has focused on estimating Origin-Destination (OD) flows to reconstruct OD matrices, either at urban or nationwide scales. OD flows differ from path flows in that they do not indicate the route taken between origin and destination, nor characterize the distribution of the OD flow between different possible paths. The most significant related works have already been presented in Chapter 2.1.1 [Nanni et al., 2014, Çolak et al., 2015, Alexander et al., 2015, Toole et al., 2015, Lwin et al., 2018]. These methods rely on the detection of the static phases out of the users' records. The extraction of these static phases supports estimating users' activity chains Jiang et al. [2017]. Within such an activity chain, each couple of consecutive static phases characterizes an origin destination couple along which a trip occurs. Systematically summing up the trips occurring between the same origin and destination provides an estimation of the OD matrices. Except for the work of Toole et al. [2015], which we will discuss below, the majority of these works do not address the route taken by users along their origin-destination trip. But these works still provide interesting results and a methodological framework on which to build more systematic path-flow estimates. Indeed, this additional information on routes is essential for the estimation of air emissions since these require to relate the traffic volume to the speed and thus to spatialize these volumes on the network.

However, apart from this first limitation, these works have, in our opinion, another serious limitation: they do not take into account the biases of the CDR data [Ranjana et al., 2012, Hoteit et al., 2017, Chen et al., 2019, Zhao et al., 2021]. By nature, such data are sensitive to user activities, and they do not render the users' mobility to its full extent. In particular, some stays might go undetected if users do not engage in communication activities meanwhile. This sensibility can lead to a misinterpretation of the trips performed,

shifting these trips to erroneous ODs and underestimating the overall flow or traveled distances. Several works have recently focused on characterizing the biases related to using CDR data for mobility inference. Hoteit et al. [2017] and Chen et al. [2018] propose to downsample GPS data according to CDR communication rates and compare some strategic mobility indicators in the downsampled and complete dataset. They demonstrate that a radius of gyrations, and the most significant locations can be extracted from the mobility of CDR user. However, they also show that a significant share of the locations visited by users disappear in the data downsampling, suggesting CDR data are not adapted to short term mobility analyses. From another perspective, Zhao et al. [2021] propose a method to infer whether *hidden visits* exist behind observed CDR trips. They found that 10% of the trips observed from CDR data were erroneous and missed a latent stay, suggesting the impact of neglecting these stays on the magnitude and distribution of origin-destination flows. Chen et al. [2019] propose a method to address the CDR data sparsity at a daily scale, by enriching hourly time slots with positional information learned from the users' regularity and the context. They show how completing a user's position data can have a significant impact on the results of typical mobility analyses, such as individual mobility laws or trajectory uniqueness.

The user-activity-dependent characteristic of the data also impacts the detection of the paths taken by users during their trips. In introduction of this part, we defined pass-by points. Their sparsity during trips explains that the flow-centered literature has focused on OD flows rather than path flows. The work of Toole et al. [2015] illustrates this. After estimating an OD matrix, the authors would instead use a dynamic traffic assignment method to distribute the OD flows on the road network than relying on the available pass-by points to infer the undertaken trajectories. On the other hand, this question of trip reconstruction has been the subject of extensive literature, until now somewhat disconnected from the question of flow estimation. This subject has been explored with various mobile phone data sources. The methods adopted can go from shortest path completion (Paipuri et al. [2020] with LBNS data) to machine learning methods. Based on signaling data (richer than CDR) and using Markov models, Pourmoradnasseri et al. [2019] propose a trajectory reconstruction method at a national scale (Estonia) to infer the complete base-station-scale path. The models include both individual and global features to integrate the user habits with inter-user trends. Still with signaling data, Asgari et al. [2016] proposed a trajectory reconstruction method in multi-modal networks, while Bonnetain et al. [2021] develop such a method based on the individual historical trajectories clustering. Although most of those works rely on richer temporal data than CDR data, they provide interesting options for trip completion and subsequent path flow estimations). Unsurprisingly, given the characteristics of CDR data, few attempts have been made to reconstruct path flow distribution from those [Forghani et al., 2020].

In this chapter, we propose a methodological step-wise workflow to bridge the gap between the trip reconstruction methods and path flow estimation, while taking into account the bias to the incomplete nature of the data. We adopt an activity-chain-based point of view, *i.e.*, a day of mobility is described as a succession of static phases and trips. Following some of the literature works cited above, we suggest resorting to the users' mobility regularity by exploiting historical knowledge to enrich the observed mobility. The approach we develop specifically targets users identified as locals (either residents or commuters, *i.e.*, who have enough historical) and with regular mobility, based on the entropy measurement we described in the introduction of this part. In a first step, at the daily scale, we propose a heuristical approach to enrich the activity chains extracted from CDR data, based on identifying mobility routines within the richest days of history. In a second step, at the trip scale, we propose an integrated method to map-match a subset of trips to infer the path-flow distribution and estimate the overall flows.

The contributions of this chapter are the following:

- First, we propose an activity chain completion method based on a heuristical approach. This involves the extraction of daily activity chains from the user history, the identification of users' mobility routines, and completing a sparse day of data according to this history.
- Second, we propose a simple map-matching method based on an exogenous knowledge of the paths taken at the city scale, the estimation of prevailing paths described in Chapter 2. This knowledge is complete by the users' data and used for inferring the complete paths of users displaying *pass-by points*.
- Finally, we propose a method for estimating path flow distributions based on varying estimation horizons. It allows us to distribute users who could not be map-matched onto the different paths.

## 3.2 Method overview

This section presents the methodological workflow we propose for the individual data enrichment. This workflow is illustrated in Figure 3.1. It specifically focuses on regular individuals, as defined in the introduction preceding this chapter. The mobility reconstruction methodology that we are developing is based on two aspects.

First, it consists in reconstructing the users' activity chains. This step relies on the availability of a CDR data history, with unchanging user IDs. The preliminary offline analysis of this historical data is supposed to produce individual mobility references that will serve as an analytical framework when processing and integrating new CDR data. By individual reference, we mean more specifically an individual mobility routine, which will allow to interpret and enrich the daily mobility of a user observed outside the historical framework. This is defined in the next section. Here, we artificially consider the January and February 2020 data to correspond to this history, while the March data is treated as online data. Although the data in our case study was collected in sufficient proportion from the outset, we propose to arbitrarily distinguish January and February 2020 as historical data, while the following month's data is considered as day-to-day received data. On the basis of the mobility routines constructed from the historical data, we propose to seek to enrich the activity chains of non-historical days, when necessary. Section 3.3 develops this method.

Following this reconstruction, the second aspect of the enrichment of individual mobility is the association, for each trip induced by an activity chain, with a complete trajectory at the regional scale. Then, we focus on the map-matching of the trajectories. This step relies on a preliminary enrichment of the prevailing paths extracted from the systematic network analysis (cf. Chapter 2) with the alternative significant paths extracted from the CDR observations. It allows generating an extensive paths dataset. This dataset provides a valuable insight for interpreting the users pass-by points and map-match a trip with an existing pre-identified path. This map-matching approach is easy-to-implement and cost-efficient. However, it allows only to map-match users that display pass-by points during their trips. Therefore, we also propose to use this trips subset to estimate the dynamic path flow distribution. This distribution will further be applied to distribute the remaining trips over the network and estimate overall sample flows. In a last step, the estimated flows are up-scaled to be representative of the overall population. This overall process is detailed further in Section 3.4.

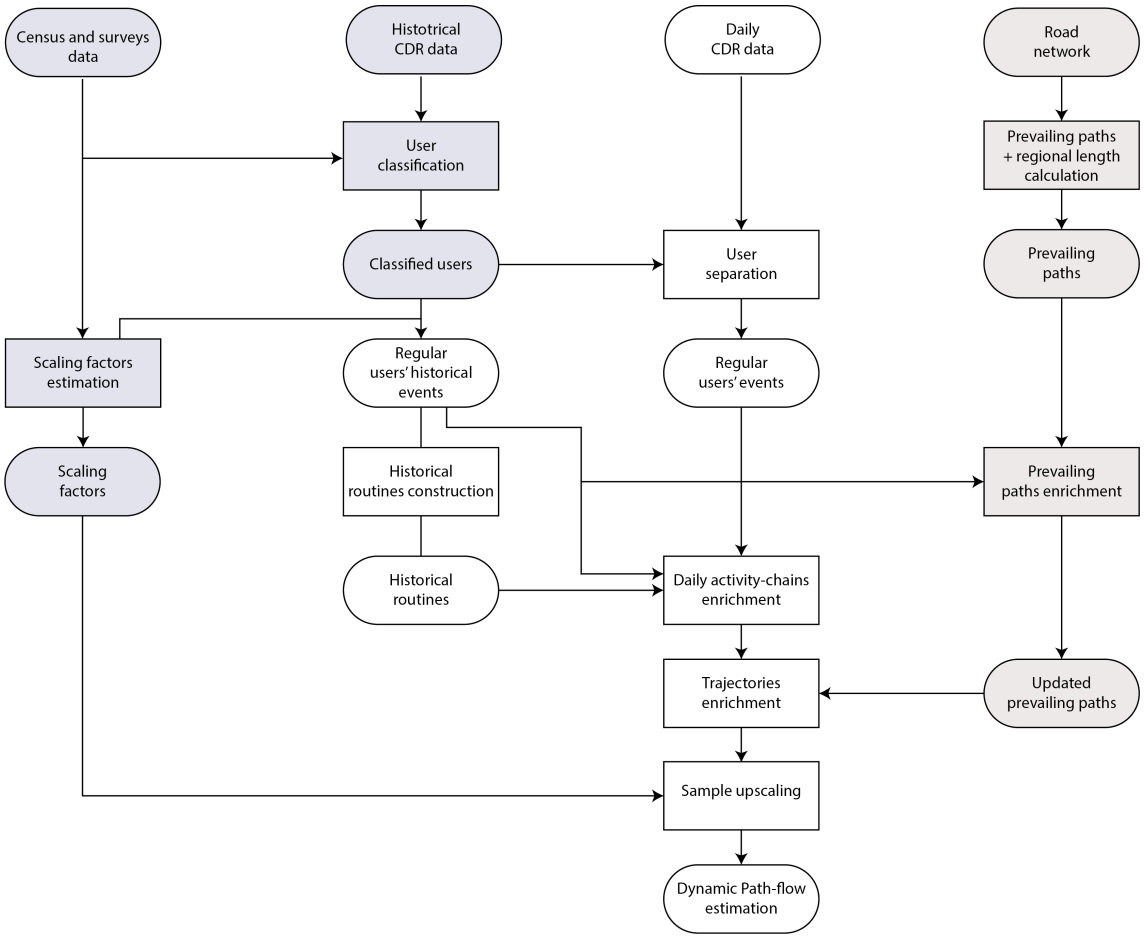



Figure 3.1: General individual mobility reconstruction workflow

## 3.3 Daily-Activity Chain Enrichment

### 3.3.1 Definitions

When analyzing a single day of CDR data for a given user, it is very likely that some stays go undetected, resulting in a partial mobility observation. However, the literature has shown that humans are mostly regular in their mobility patterns. Several works have already made use of this regularity feature and resort to available historical data in order to enrich the CDR users mobility footprints, either at the trajectory or at the daily scale. Here, we propose to adopt a similar strategy to enrich the users daily mobility based on their historical analysis. On the contrary to Chen et al. [2019], who adopt a discrete (hourly) representation of the daily mobility, we propose to stick with an activity-based daily representation [Axhausen and Gärling, 1992, Chu et al., 2012], *i.e.*, a representation of the daily mobility as a succession of activities located in time and space. With mobile phone data, Jiang et al. [2017] proposed an interesting work with this mobility conception. It allows to keep a clear identification of the activity beginnings and ends and hence identify the trips timestamps. In the context of CDR data use, we define:

**Definition 6 (Activity)** *An activity $a$ of a user $u$ is a period of time during which a user is identified at a specific position. Let $a^u = (l, t_{start}, t_{end})$, where $l$ is the stay location, $t_{start}$ is the timestamp of the beginning of the stay, and corresponds to its first communication event, and $t_{end}$ is the timestamp of the end of the stay, and corresponds to its last communication event.*

**Definition 7 (Activity chain)** *An activity chain $c^u$ of a user $u$ is the daily succession of activities of a user. Let $c^u = (a_1^u, a_2^u, ..., a_n^u)$. An activity chain can be an incomplete representation of the user's daily mobility, considering that some stays may have been undetected.*
*We further call a **spatial activity chain** $c_t^u$ of a user $u$ the ordered vector of the locations visited during the day: $c_t^u = (l_1, l_2, ..., l_n)$.*

**Definition 8 (Sequential completeness)** *We define the sequential completeness metric $\rho_c$ as the fraction of the day covered by sequence data:*

$$\rho_c = \frac{\sum_s s.t_{end} - s.t_{start}}{24} \tag{3.1}$$

*where $s$ is a sequence of events of the considered user, and $s.t_{start}$ and $s.t_{end}$ are respectively the beginning and ending timestamps of the sequence. The duration of this sequence is expressed in hours and related to the number of hours in a day to evaluate the daily data coverage. We call **complete activity chain** any activity chain $c^u$ characterized by a minimal sequential completeness $\rho_C$, set to 0.8 in this study. Therefore, any activity chain characterized by a lower completeness will be considered as incomplete, while any activity chain with a higher completeness will be considered as complete.*

**Definition 9 (Mobility routine)** *A mobility routine $r^u$ of a user $u$ is defined as a cluster of similar and complete activity chains $c^u$. Each cluster is further associated with its spatial represented, which we define as the spatial activity chain of the most frequent chain of the routine.*

**Definition 10 (Mobility profile)** *A mobility profile $p^u$ of a user is defined as the set of routines identified in one user's historical mobility, in which each routine is associated with its weight compared to other routines. Depending of the user's regularity, it might be made of few to several different routines. If the user's mobility has absolutely no regularity, then the profile might as well be empty of any routine.*

With respect to data completeness, several works in the literature base their analysis on a discretization of time, where the completeness rate is defined as the share of time slots during which a communication event takes place. However, such a metric is unavailable in our case because the data has a specific compressed sequence format (cf. Chapter 1), in which the intermediate timestamp information has been erased. Due to the sequence format of the data, a long sequence may be observed from a location, while the user could have only generated there only a few events, leaving space for movements between the measured events. However, the timestamps of these events being unknown, these sequences cannot be further enriched. In the future, access to less compressed data will allow to ignore this problem.

### 3.3.2 Routines Construction

We propose to proceed to a clustering of complete activity chains based on their spatial dimension only. The first objective of this approach is to introduce an important flexibility regarding the distortion of activity chains from one day to another. While the spatial activity chains may be quite repetitive, the temporal pattern may vary from one day to another, due to work constraints, congestion or transportation constraints. The objective is also to build routines that are more provided than if the temporal dimension was taken into account in the clustering.

The distance metric used to compare the daily spatial activity chains of a user is based on the Longest Common Subsequence (LCS) research. The LCS research is a problem frequently associated with text mining and natural language processing. The objective is to determine the longest subsequence of elements (characters for instance) between different sequences of such elements. Derived from this analysis, the length of the longest common subsequence $l_{LCS}$ can be considered as a measure of similarity between two elements.

We propose to apply the LCS problem to measure the distance between activity chains. The spatial activity chains correspond to the sequences in which we look for the longest common spatial subsequence. Because of the potential signal balances and echoes, we set a distance tolerance threshold, so that two locations (or base stations) can be considered as equal if their euclidean distance is less than one kilometer. Below, we provide an example of the LCS result when comparing two activity chains or words $ABC$ and $ADC$.

$$l_{LCS}(ABC, ADC) = \begin{cases} 3 & \text{if } d_E(B, D) < t_D \\ 2 & \text{if } d_E(B, D) \geq t_D \end{cases} \tag{3.2}$$

Based on this longest common subsequence similarity metric, we define a distance metric as follows:

$$d(c_1^u, c_2^u) = 1 - \frac{l_{LCS}(c_1^u, c_2^u)}{max(len(c_1^u), len(c_2^u))} \tag{3.3}$$

First, we normalize the similarity metric with the length of the longest activity chain. This normalization highlights the relative importance of the intersection between the two activity chains. Then the transformation of the similarity through the affine function $f(x) = 1 - x$ allows to turn the similarity metric into a distance metric.

Based on this metric, we compute, for each user, the distance matrix based on the pair-wise distance evaluation between the different daily complete activity chains. Then, a DBSCAN clustering method [Ester et al., 1996] is applied to the distance matrix to compute clusters of close activity chains. We select the following clustering parameters. The minimum cluster size is set to 1, which means that we do not eliminate outliers: any

observed and sufficiently rich activity chain is considered as trustworthy. As input chains are considered completed, the minimal distance between two chains should be relatively high in order to cluster them. The $\epsilon$ parameter (minimal distance for two activity chains to be aggregated in the same cluster) is set to 0.5. As the computational time for comparing long activity chains can be significant, any chain displaying more than 8 activities during a day are directly considered as a cluster, without trying to reconcile it with another activity chain. In the end, for each cluster, the spatial mobility routine is defined as the most frequent spatial activity chain observed within the cluster.

The process was applied to a subsample of 43,168 regular users. Among them, more than 35,807 were assigned a mobility profile, which is about 83% of this sample. This percentage drops to 79 if we do not include the default clusters, i.e. all activity chains with more than 8 activities. This population has an average of 4 clusters, and the average size of these clusters is 6 days of data. These numbers are promising because they indicate that a large portion of the population now has a reference on the basis of which to interpret and complete a new daily activity chain. In future work, we would like to implement a more flexible completeness threshold. For example, rather than being evaluated in relation to a full day (24 hours), this level of completion could be measured in relation to the users' own maximum daily coverage. The objective would then be to complete a user's mobility as best as possible, according to their own activity levels. This would make it possible to build mobility profiles for the entire regular population, regardless of their communication habits.

### 3.3.3 D-day mobility enrichment

This section targets the in-line integration and processing of new data. Let's consider a data day $d$ outside the historical period. Regular users $\mathscr{U}_d^r$ are identified within the incoming data (they represent approximately 62% of the daily users) and their daily activity chains are extracted. Again, we use the completeness criterion to identify the users targeted by the mobility enrichment. For a given user, if their completeness rate measured on day $d$ is higher than 0.8, the activity chain is again considered as complete and will not require enrichment. On the other hand, if the data completeness rate is below this threshold, we will try to complete the daily mobility by using mobility profiles.

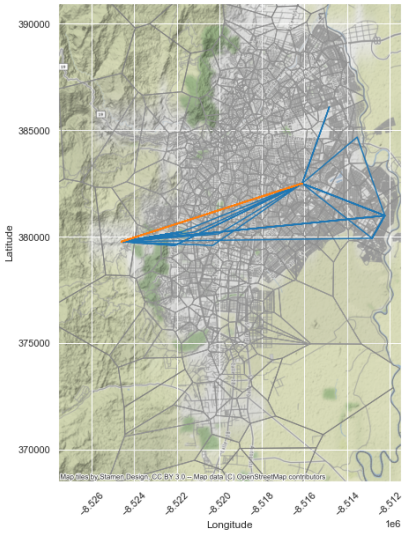
For such an activity chain to be completed, it also requires the user to have an identified mobility profile. Other constraints, that will be further presented in this section, progressively limit the number of activity chain that need or can actually be completed. For instance, users with incomplete daily activity chain and with a historical mobility profile represent 36% of the daily regular users. We will further refer to this set as $\mathscr{U}_d^{r,0}$ Any regular user whose activity chain could not be enriched on a day $d$ will have their mobility re-integrated in the mobility enrichment process from Section 3.4.

Let $\mathscr{C}_d^0$ be the set of activity chains of the users $\mathscr{U}_d^0$. We propose an heuristical approach for completing the activity chains $\mathscr{C}_d^0$. For each user $u$, the method takes in the daily observed activity chain $c_d^u$ and the user's mobility routine and returns either an enriched activity chain, if possible. The workflow follows the following steps:

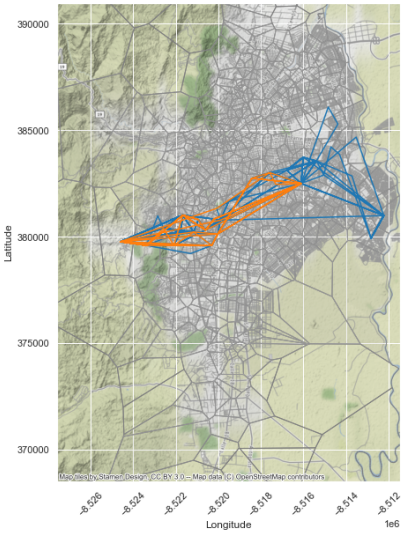
1. The incomplete daily activity chain $c_d^u$ is related to the user's mobility profile. Within this cluster, all the mobility routines for which the daily activity chain is a subsequence are considered as a candidate model activity chain.

2. Based on the weight of each routine in the profile, we sample the model activity chain from the candidates.

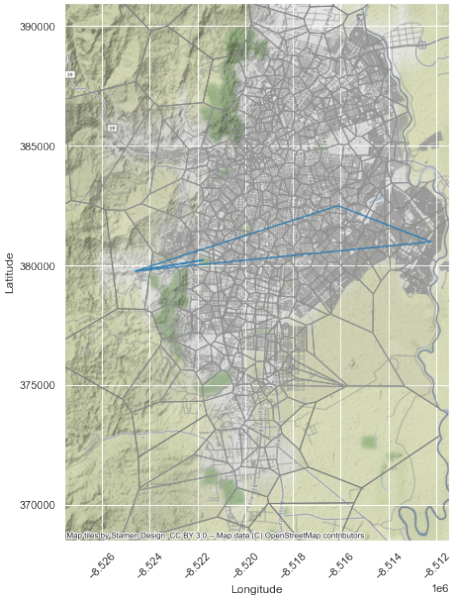
3. Comparing the model activity chain with the daily activity chain allows to identify the missing activities that should be filled.



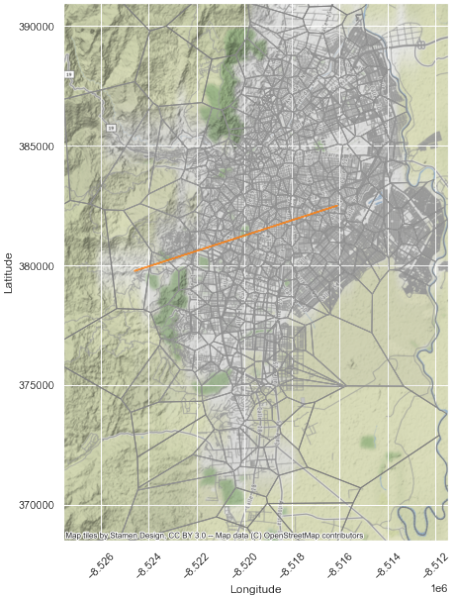

(a) Clustered activity chains



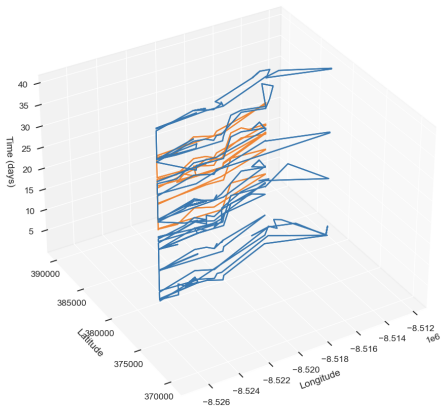

(b) Overall CDR traces




(c) Routine #1




(d) Routine #2




(e) 3D visualization of the CDR footprints

Figure 3.2: Geographic visualization of the routine identification from one user's historical CDR footprints.

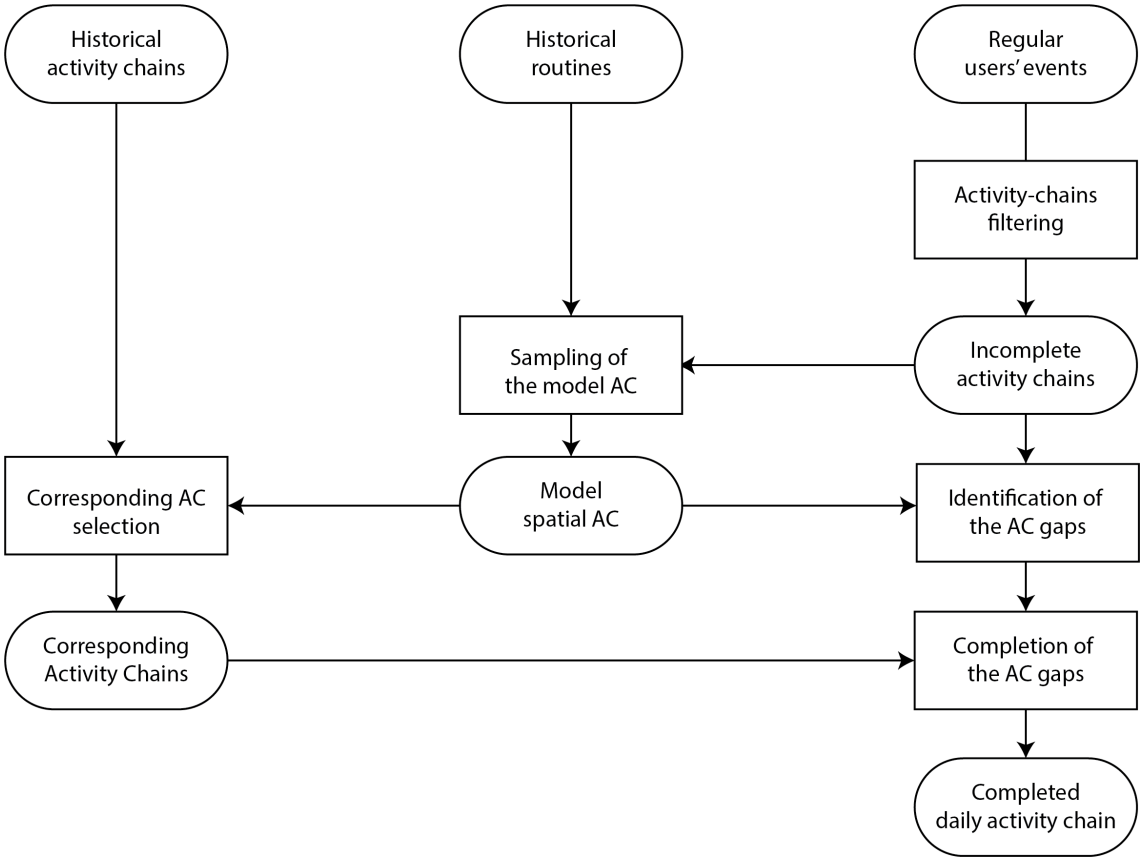



Figure 3.3: Activity chain completion workflow

4. Simultaneously, browsing the historical data provides the activity chains corresponding to the model activity chain.

5. These historical activity chains support the completions of the gaps within the daily activity chain.

This workflow is illustrated in Figure 3.3.

We further detail the gap completion process. First the question is of identifying the gaps within the daily activity chain that will need to be completed. Let $c_s^{ref}$ be the model spatial chain we want to enrich an observed daily activity chain $c^d$ with. Let $c_s^d$ the spatial chain associated with $c^d$. The spatial chain $c_s^d$ is a sub-sequence of $c_s^{ref}$. An element-wise comparison of the two sequences can return the locations of $c_s^{ref}$ that are missing in $c_s^d$, along with the position they need to be introduced at.

Then, for each identified gap in the daily activity chain, we are going to try to fill it using historical data. Formally, we define a gap of $c^d$ a couple of consecutive activities of $c^d$ that are not consecutive in the model chain $c_s^{ref}$:

$$g_{kl} = (a_k, a_l) \qquad |k+1 < l \tag{3.4}$$

Let $d_{kl}^d$ be the observed gap duration in the daily activity chain:

$$d_{kl} = a_l.t_{start} - a_k.t_{end} \tag{3.5}$$

Let $\mathscr{C}_{h,u}^{ref}$ the set of activity chains corresponding to the spatial chain $c_s^{ref}$. Let $c^i = (a_j^i)$ any historical activity chain in $\mathscr{C}_{h,u}^{ref}$. Let $c_{kl}^i = (a_j^i)_{j \in ]k,l[}$ the sub-sequence of $c^i$ in between

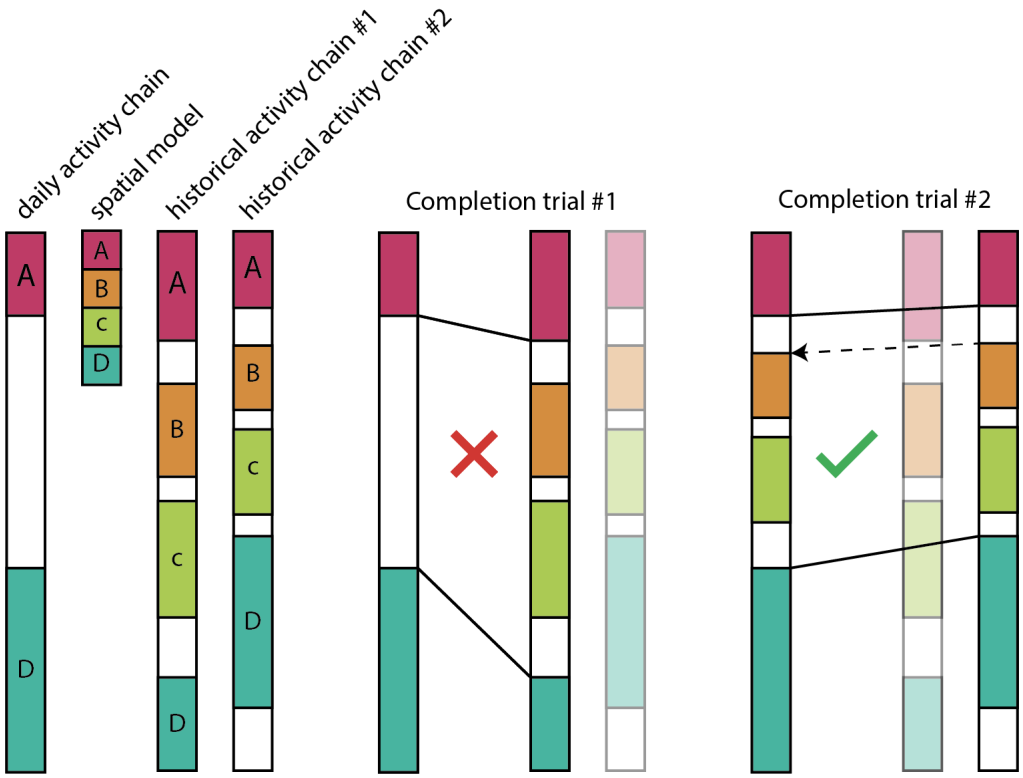



Figure 3.4: Gap completion process

the activities $a_k^i$ and $a_l^i$. And again, let $d_{kl}^i = a_l^i.t_{start} - a_k^i.t_{end}$ the corresponding duration in $c^i$. If $d_{kl}^i \leq d_{kl}$, then the sub-sequence $c_{kl}^i$ is a candidate for completing the gap $g_{kl}$.

For each gap $g_{kl}$, we sample among the candidate completion sequences. The selected completion sub-sequence $c_{kl}^i$ is translated in order for the inter-activity times observed in $c^i$ to be respected in the enriched version of $c_d$. Figure 3.4 illustrates the selection of the completion sub-sequence and the adaptation of the activity times according to the time constraints imposed by the daily activity chain.

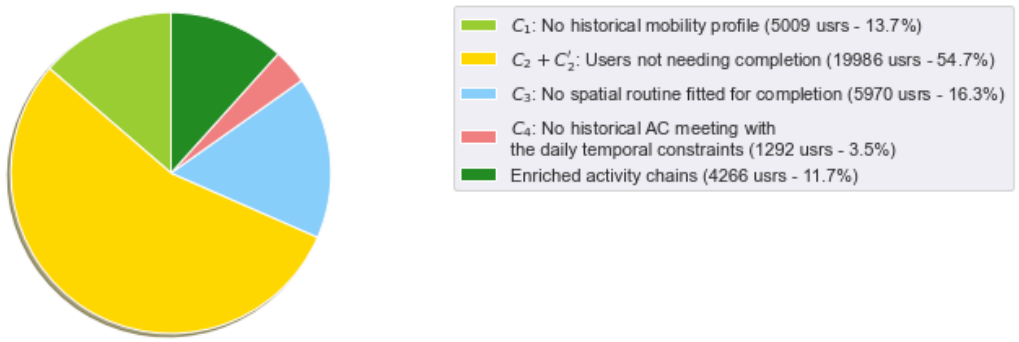



Figure 3.5: Shares of the daily regular population according as processed by the activity chain enrichment workflow

### 3.3.4 Results

In this section, we assess the impact of the activity chain completion on the origin-destination matrices and display the resulting origin-destination matrices, which we compare to the available ground-truth data.

#### 3.3.4.1 Impact assessment

Among the regular users observed on the day $d$, the completion process targets those with an historical mobility profile (condition $C_1$) and whose daily mobility is incomplete (condition $C_2$). Yet, this does not mean that all the remaining users will have their mobility enriched. In fact, for the daily mobility to be enriched with historical activities, the following conditions must be met:

- $C_3$: an historical routine should corresponding to the daily activity chain;
- $C_2'$: the selected routine should be different from the daily activity chain, otherwise the daily data is again considered as not requiring enrichment;
- $C_4$: the temporal constraints of the daily activity chain should be compatible with some historical routines for the daily gaps to be completed.

There is therefore effective completion of individuals only when these different conditions are met. In practice, on a given day of data this represents a limited part of the population of regular individuals. Figure 3.5 illustrates how each of these conditions contributes to restricting the number of users candidates for the enrichment process. On a daily sample of 43,168 regular users, about 11% of them see their mobility effectively enriched.

In fine, the completion of the mobility of these users results in a multiplication by a factor of 2.3 of the number of trips observed. However, it must be qualified with respect to the results observed in Figure 3.6. This figure shows a good reproduction of the morning an afternoon peaks and of the daily dynamics, and reasonably increase the magnitude of the number of trips during the day. Yet, we observe a significant increase in the number of trips observed at night-time, and on Sunday. We believe this observation results from different biases related to our method.

1. First, this anomalous mobility reconstruction is observed during periods of lower communication and mobility activity. Few activity chains should be able to enrich these periods of the day or week, but those that are may be over-weighted and result in an over-enrichment of the daily activity chains on their model. In particular,

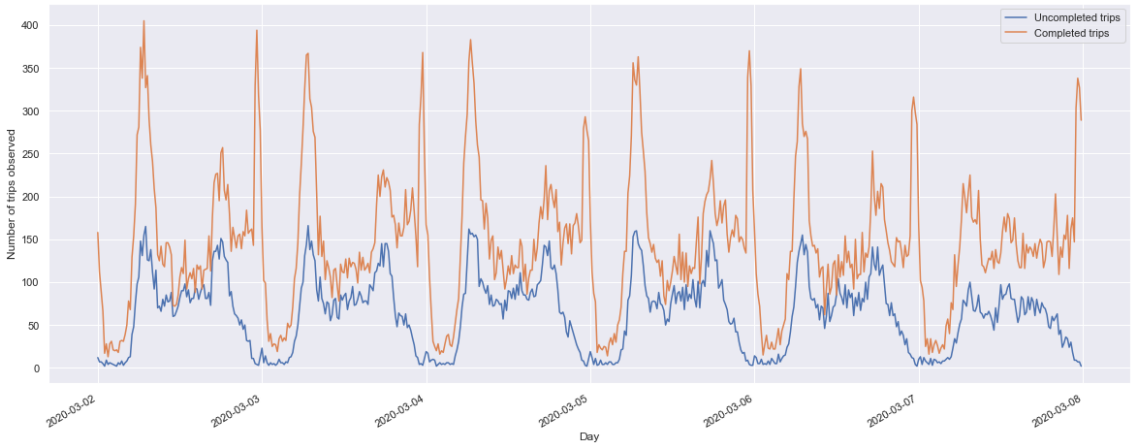



Figure 3.6: Comparison of the number of traveling users over a week of data, with and without activity chain completion, for users whose activity chains could be enriched

commuters are likely to be particularly affected by this problem. While they most likely are absent during night time, with daily coverage below 0.8, any routine during which they are observed at night in the city can be considered as an enrichment reference. This example highlights the need to differentiate at least distinct completeness thresholds for different user categories.

2. For reasons of computation time optimization, we have declared not to aggregate the longest activity chains in cluster. These chains will therefore constitute mobility routines on their own, that are likely less reliable than the routines resulting from the clustering process. These routines can also explain, in a certain dimension, an excessive reconstruction of mobility. In future work, it will be possible to optimize the aggregation process or to aggregate these routines.

3. Finally, the very excessive reconstruction of mobility on Sundays may be related to the non-distinction, in our framework, of the workday and weekend activity chains, which are probably distinct. This distinction seems relevant to introduce in view of the results.

These are points of vigilance to be taken into account in future improvements of the method, and raise in particular the question of validation, which we discuss in Section 3.5.

As it stands, this enrichment of the daily activity chains contributes to a 11% increase in total mobility. This suggests a significant impact on estimated regional flows and traffic volumes. Figure 3.7 shows the impact of this completion on the number of trips during the first week of March.

In line with these results, a number of studies seem relevant and are envisaged as future work to:

1. characterize the biases and strengths of the enrichment method;

2. make it more robust and less sensitive to the own biases of the data (temporal availability, signal balance, ...);

3. evaluate the impact of this enrichment on mobility.

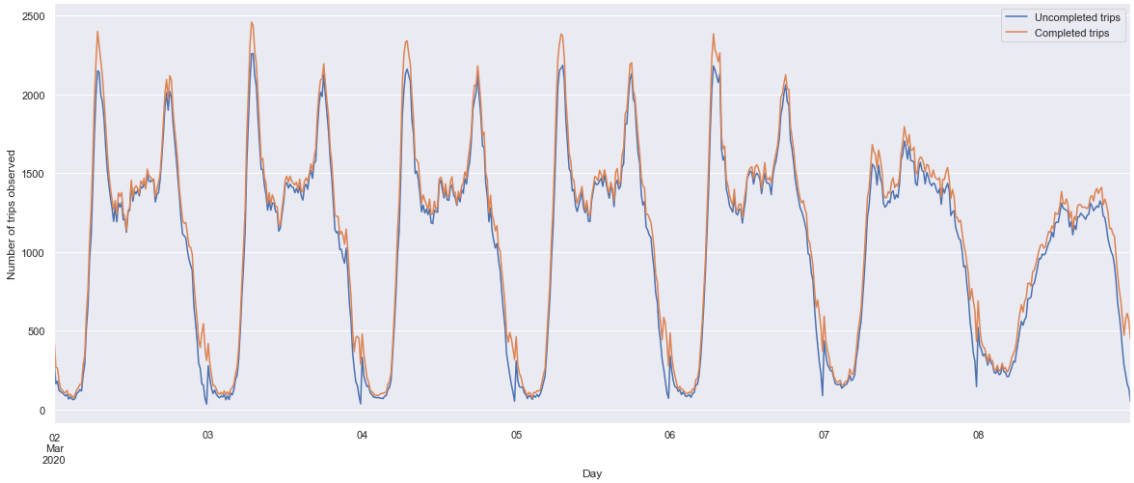



Figure 3.7: Comparison of the number of traveling users over a week of data, with and without activity chain completion, for regular users


(a) Reference daily OD matrix

(b) Completed daily OD matrix

Figure 3.8: Reference and observed origin-destination matrices at the *comuna* scale

#### 3.3.4.2 Comparison with available ground truth data

In this section, we propose to compare the destination matrix derived from the CDR data (after population scaling) with the available ground truth data. These data correspond to the mobility survey conducted in Cali in 2015 [Metro Cali, 2015].

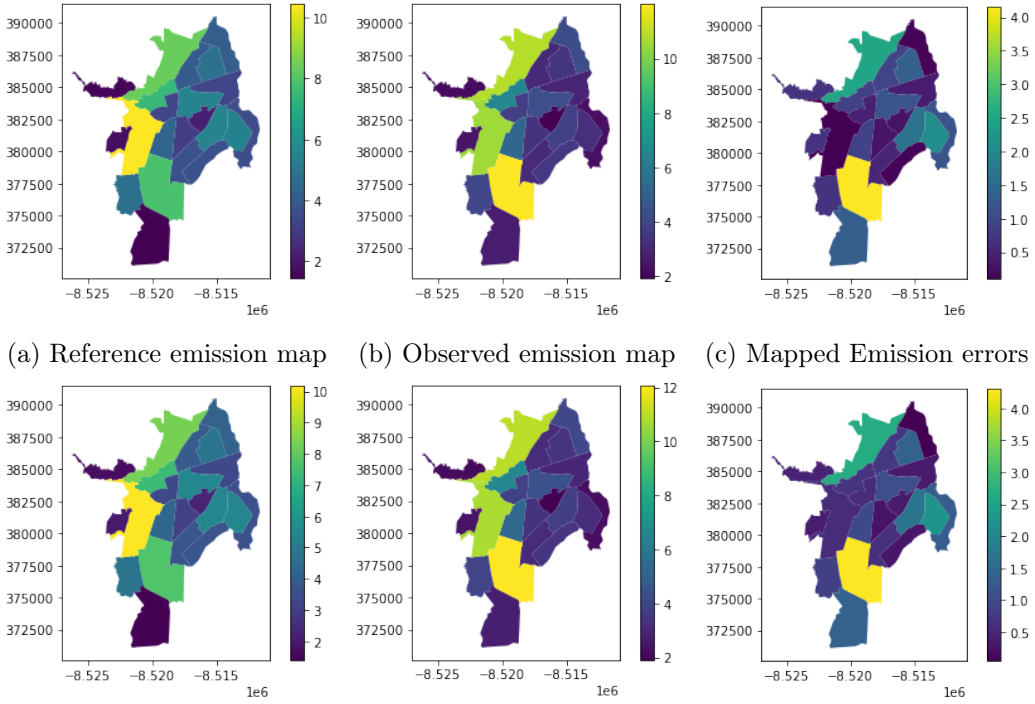
Figure 3.8 provides a comparison of the reference origin-destination matrix (left) with the origin-destination matrix derived from the completed CDR data for the 2$^{nd}$ of March 2020. We normalize the two matrices for comparison, the origin destination matrix from Cali being based estimated on commuting trips. We observe that the general structure of the matrix is preserved. In particular, the north-south structuring axis is well identified. However, there is also a relative underestimation of the flows to and from the east of the city.

We also measure the numerical error between these two normalized matrices according to two metrics: the mean absolute error (MAE) and the root mean squared error (RMSE), in percents of the total daily flow. These results are compared with the same results when computed on a raw non-completed origin-destination matrix. The results translate a small reduction of the errors when working with the completed matrix.

Figure 3.9 provides a comparison of the reference emission and attraction vectors with

| | MAE | RMSE |
|---|---|---|
| Enriched ODM | 0.116 | 0.259 |
| Raw ODM | 0.117 | 0.265 |

Table 3.1: Comparison of the MAE and RMSE errors between the raw and enriched origin destination matrices




(a) Reference emission map (b) Observed emission map (c) Mapped Emission errors

(d) Reference attraction map (e) Observed attraction map (f) Mapped attraction errors

Figure 3.9: (a), (b), (d), (e): Reference and observed origin-destination matrices at the urban district scale, where the color gradient refers to the weight of the region in the global emissions/attractions. (c), (f): Squared errors of the region weights in the emission and attraction vectors

the ones derived from the CDR enriched data. Emission and attraction vectors record the contribution of each district to the total trips emitted or received during the day. Here again they were normalized by the total trip amount. It also displays a mapping of the errors for both the emission and attraction vectors. Comparing the reference and observed emission and attraction maps confirms the similarity of the structure of the mobility. However, this also highlights the bias on areas east of the city center, whose participation in mobility is underestimated. This also emphasizes the overestimation of mobility to and from the *Comuna* 17 in the South of Cali (in yellow in Figures 3.9c and 3.9f).

## 3.4 Trip Enrichment

Once the activity chains have been completed, the challenge is to spatially reconstruct the trips between these different activities. This is the purpose of this section.

### 3.4.1 Definitions

Let $\mathscr{R} = \{r \in \mathscr{R}\}$ the regional partitioning of the network and let $p = (r_1, ..., r_n)$ a regional path in $\mathscr{P}$. We introduce the following definitions.

**Definition 11 (Observed path)** *We define an observed path $p_{obs}$ as the sequence of regions within which pass-by points (Definition 5) were observed between two consecutive static phases. Assuming at this stage that no in-between static phase is missing, $p_obs$ is a sub-sequence of the real path $p$ traveled by the user from the origin static position to the destination one.*

**Definition 12 (Trip)** *A trip $\tau$ is defined as a ternary structure $(p, t_{start}, T)$ relating a regional path $p$, a trip starting timestamp $t_{start}$ and a trip duration $T$. A trip characterizes the movement of an individual on the network on their way from their regional origin $O$ to their regional destination $D$.*

**Definition 13 (Observed trip)** *By extension, we define an observed trip as a trip extracted from an activity chain (Definition 7)). Considering $c = (a_1, ..., a_n)$ and two consecutive activities $(a_i, a_{i+1})$ of $c$, let:*

- *$p_{obs}$ be the observed path between $a_i$ and $a_{i+1}$;*
- *and let $T_{obs}$ be the observed duration between $a_i$ and $a_{i+1}$: $T_{obs} = a_{i+1}.t_{start} - a_i.t_{end}$.*

*The observed trip between $a_i$ and $a_{i+1}$ is $(p_{obs}, a_i.t_{end}, T_{obs})$*

These definitions highlight the differences existing between real trips traveled by users and observed trips as extracted from the CDR data generated before, during and after the trip. In this section, we intend to address the issue of the parcel nature of the regional path extracted from the CDR data $p_{obs}$. The other key issue in this extraction of trip data, which is the time bias of these data included in $T_{obs}$, will be the subject of Chapter 5.

### 3.4.2 Prevailing paths enrichment

The prevailing path detection method, described in Chapter 1, provides a set of regional paths $\mathscr{P}_0$ that characterizes the most probable paths to travel from a regional origin to a regional destination. Thereafter, $\mathscr{P}_0^{OD}$ refers to the subset of prevailing paths from $\mathscr{P}_0$ between the regional origin $O$ and the regional destination $D$.

Because the prevailing path estimation method is based on the calculation of the shortest path in distance, it likely misses the most significant detours made by the population to by-pass crowded region, for instance in peak hours. Therefore, we suggest completing the regional path set $\mathscr{P}_0$ with additional alternative paths derived from the CDR data observations.

To do so, we extract from the historical data all continuous trips, *i.e.,* all observed trips whose regional path is made of adjacent consecutive regions. Let $\mathscr{T}_h^{OD,cont}$ be the set of historical continuous trips traveling from the regional origin $O$ to the regional destination $D$. We name $\mathscr{P}_{obs}^{OD,cont}$ the corresponding set of continuous regional paths:

$$\mathscr{P}_{obs}^{OD,cont} = \{\tau.p | \tau \in \mathscr{T}_h^{OD,cont}\} \tag{3.6}$$

This set of continuous paths is used to extend $\mathscr{P}_0^{OD}$ into a new set of reference regional paths $\mathscr{P}^{OD}$:

$$\mathscr{P}^{OD} = \mathscr{P}_0^{OD} \cup \mathscr{P}_{obs}^{OD,cont} \tag{3.7}$$

By extension, the overall set of regional paths $P$ is built as:

$$\mathscr{P} = \mathscr{P}_0 \cup \mathscr{P}_{obs}^{cont} = \cup_{OD} \mathscr{P}^{OD} \tag{3.8}$$

In practice, in order to limit the size of $\mathscr{P}_{obs}^{cont}$, we establish it based only on the continuous CDR trips that display limited intersection with the known prevailing paths $P_0$. Therefore, the objective is mostly to identify alternative paths to the shortest paths. We also target trips shorter than one hour, so as to avoid incorporating anomalies and excessive detours in the considered path. Finally, we consider short-distance trips (less than one kilometer) are less prone to detours than longer ones. Therefore, for close origins and destinations, the prevailing paths from $\mathscr{P}_0^{OD}$ should represent well the possible paths, while additional information learned from CDR-based continuous paths might be significantly affected by signal balance and echoes. In the end, the trips selected out of the historical period represent 1.27% of the overall trip set.

We observed that this enhancement process resulted in multiplying by 1.65 the number of regional paths, and on average, by multiplying by 1.98 the number of prevailing path for each OD couple. This indicates that on average one alternative path per origin-destination pair is detected. In order to target more precisely the trips considered for enrichment, one can consider focusing on paths complying with a predefined azimuth range, as it was done in Chapter 2, or putting a maximum limit on the path detour ratio. These approaches could be implemented as part of the improvement of this method.

Figure 3.10 gives an example of the added value of including CDR-based path into the original prevailing path database. While the paths detected according to the method implemented in Chapter 2 follow a simple North-South axis through the city center of Cali (left map), an alternative by-passing the city center via the 25th Street and the Simon Bolivar ring road emerges from the CDR data.

Figure 3.11 gives a more general illustration of the spatial dispersion of the prevailing paths before an after the CDR-based enrichment. Based on the metric defined by Equation 2.29, we evaluate, for each OD, the intersection rate between its different prevailing paths. We clearly see that while the original prevailing path dataset (in blue) present rather high intersection level, the CDR-based one displays much lower one (in orange) and contributes to the spatial diversity of the prevailing paths when integrated to the original path set (green).

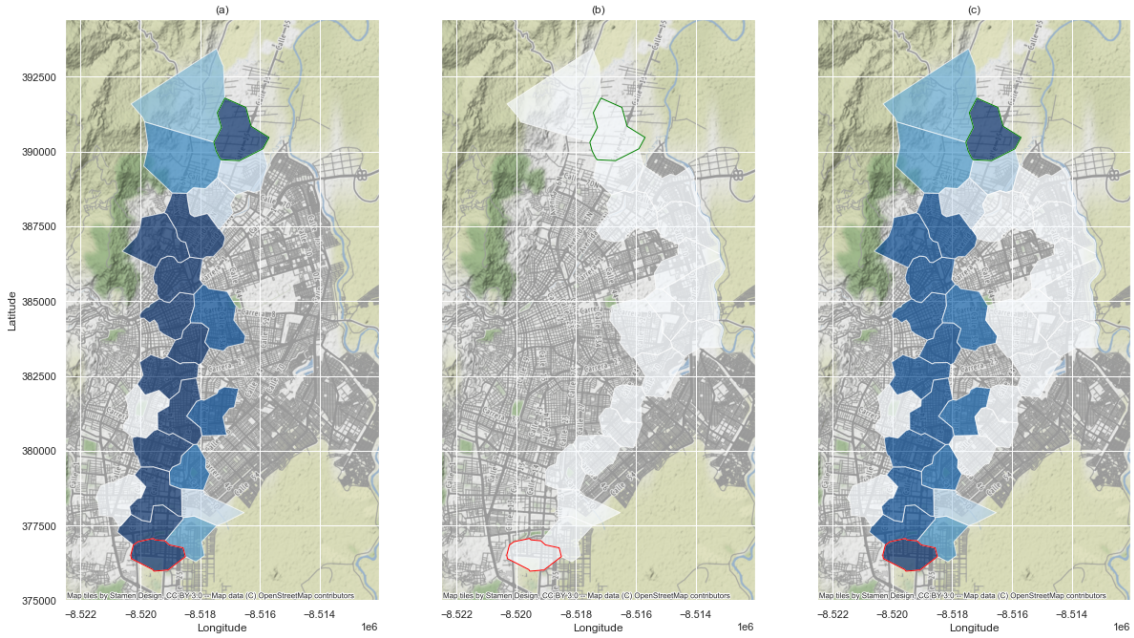



Figure 3.10: Enrichment of the prevailing paths based on the CDR observations. From left to right: (a) prevailing paths from the systematic network analysis, (b) alternative path from CDR data analysis, (c) completed prevailing paths set. The green-circled region corresponds to the origin region while the red-circled one corresponds to the destination region.

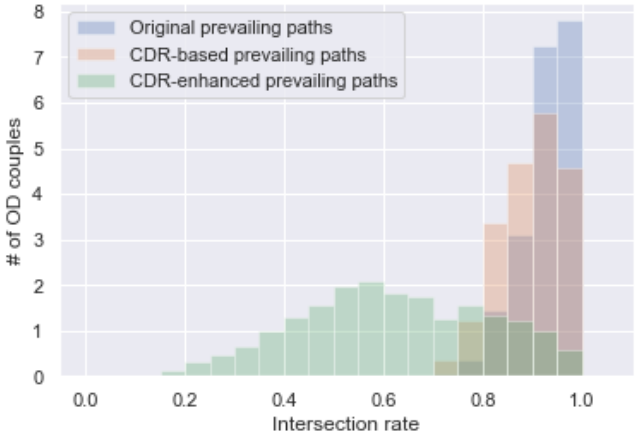



Figure 3.11: Histogram of the intersection rate distribution for the different paths set.

### 3.4.3 Trips map matching

The enhanced prevailing path database provides an insight into the urban mobility, through the lens of which we can analyze and enrich the overall CDR trips database. In particular, path flow estimation framework we propose relies on this prevailing path knowledge to map-match users and trips on the network and reconstruct the individual and sample mobility. Let $\mathscr{T}_d$ be the set of trips extracted from the daily regular users $\mathscr{U}_d^r$. We define as $\mathscr{T}_d^*$ the subset of $\mathscr{T}_d$ made of trips displaying pass-by region information, *i.e.*, with pass-by point information elsewhere than in the origin and destination regions $t.O$ and $t.D$:

$$\mathscr{T}_d^* = \{t \in \mathscr{T} | t.p_{obs} \neq \ [t.O, t.D]\} \tag{3.9}$$

This set was found to represent 37% of the observed trips. They correspond to the trips targeted for map-matching.

We additionally define a similarity metric $s_\cap$ between a trip $\tau$ from $\mathscr{T}_d^*$ and a regional path $p$ based on the measurement of the regional intersection of the path $p$ with the observed sparse path $t.p_{obs}$ of the trip $\tau$.

$$\begin{aligned} s_\cap : & \mathscr{T}_d^* \times \mathscr{P} \longrightarrow \mathbb{R}^+ \\ & (\tau, p) \longrightarrow \frac{|\{r \in \tau.p_{obs} |\ r \in p \setminus \{O, D\}\}|}{|\{r \in \tau.p_{obs}\{O, D\}\}|} \end{aligned} \tag{3.10}$$

This metric supports the implementation of a map-matching process based on the identification of the path $p$ from $\mathscr{P}$ that has the largest similarity with the observed trip $t$:

$$\begin{aligned} f_m : & \mathscr{T}_d^* \longrightarrow \mathscr{P} \\ & \tau \longrightarrow argmax_{p \in P}(s_\cap(\tau, p)) \end{aligned} \tag{3.11}$$

This allows to generate a set of spatially enriched trips $\mathscr{T}_d^{*\prime}$, to which we will refer in the next section.

Figure 3.12 illustrates the impact of the enrichment of the prevailing path set conducted in the previous section on the trip map-matching. Using the similarity metric defined in Equation 3.10, we measure in both cases the intersection of each trip with the path that intersects it the most. The plot shows that the CDR-enhanced prevailing paths set allows to better explain the mobility tracks. In particular, the mean intersection rates goes from 0.74 to 0.79, and the number of trips with an observed paths completely included in a prevailing path raises by 21% when using the CDR-enhanced data set rather than the original one.

### 3.4.4 Adaptative path flow estimation

The map-matching conducted in the previous section allows to enrich the spatial dimension of user trips when pass-by points are available. However, a significant portion of the trips $\mathscr{T}_d$ have no such intermediate points, making it impossible to infer the path taken.Yet, these trips need to be spatialized on the network for a complete estimation of path flows and regional volumes. We propose to adopt a probabilistic approach to meet this challenge, based on the estimation of path flow distribution. Path flow distribution characterizes, for a given OD couple, the dynamic distribution of the OD flows between the various potential paths $\mathscr{P}^{OD}$. In practice, for every $p$ in $\mathscr{P}^{OD}$, it is defined by a coefficient $\alpha_p^{OD}$:

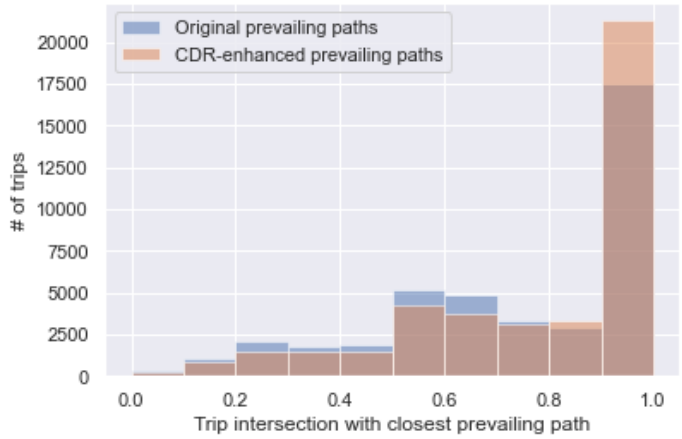



Figure 3.12: Distribution of the trip intersection with the closest prevailing path, when matching with the original prevailing path dataset (in blue) and with the CDR-enhanced prevailing path dataset.

$$\alpha_p^{OD} = \frac{N^{OD,p}}{N^{OD}} \tag{3.12}$$

where $N^{OD,p}$ is the path flow of travelers from $O$ to $D$ by path $p$, and $N^{OD} = \sum_{p \in \mathscr{P}^{OD}} N^{OD,p}$ is the OD flow of travelers from $O$ to $D$.

To further enrich the remaining trips, we propose the following method. In a first step, we propose to resort the map-matched trips $\mathscr{T}_d^{*\prime}$ to estimate the path flow distribution. In a second step, we suggest to apply these distributions to the remaining trips to map-match, *i.e.*, $\mathscr{T}_d \setminus \mathscr{T}_d^{*\prime}$.

Yet, estimating the path flow distribution from $\mathscr{T}_d^{*\prime}$ raises two related issues: 1. the temporal granularity of the calculation and 2. the statistical representativity of the data on which these coefficients are calculated. On the one hand, a fine temporal granularity makes it possible to account for coefficient variations, and thus for the redistribution of flows as a function of time and traffic conditions. On the other hand, the finer the granularity, the less the corresponding samples are provided, which raises the question of the reliability of the conclusions drawn. Moreover, there is variability between different times of the day, or different OD pairs, so that some OD pairs will be highly observed at a period $t$, while others much less so.

Therefore, we propose an adaptive path flow distribution estimation method, whose time evaluation horizon varies according to the OD pair considered and the time of the day. This method relies on the definition of a set of different temporal resolutions: (15 minutes, 30 minutes, 1 hour, 2 hours, 4 hours, 8 hours, and 24 hours). This geometric progression with the common factor 2 (up to 8 hours) ensures a simple change of scale and the switch from one granularity to another by multiplication. This sequence provides 7 different and compatible temporal scales to estimate the path flow distribution. For a given OD pair, the method we propose involves estimating the path flow distribution at the finest temporal scale that ensures an OD flow above a predefined threshold. The purpose of this flow threshold is to ensure that the estimated path flow distribution is statistically representative.

We define a matrix $M^{OD} = (m_{i,j}^{OD}) \in \mathbb{R}^{7 \times 360}$ that characterizes the dynamic OD flow at different temporal resolutions:

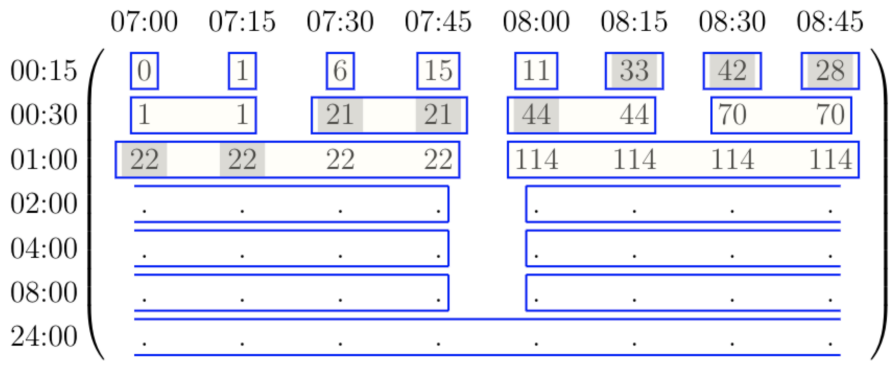



Figure 3.13: Example of $M^{OD}$ matrix for a limited time window. The blue boxes illustrate the row-by-row coefficient calculation, where the unique box value is calculated as the sum of the values in the boxes above. The numbers in gray highlight, for each time step, the temporal resolution that will support the path flow distribution estimation, for a minimal frequency threshold of 20 users.

$$m_{i,j}^{OD} = \begin{cases} N_j^{OD} & \text{if } i = 1 \\ \sum_{j \in ]n \cdot 2^{i-1}, (n+1) \cdot 2^{i-1}]} N_j^{OD} & \text{if } i \in [2, 6] \\ \sum_{j \in [1,360]} N_j^{OD} & \text{if } i = 7 \end{cases} \tag{3.13}$$

Each row corresponds to a different temporal resolution, the first row corresponding to the finest resolution (15 minutes), while the last one correspond to the coarser one (24 hour). Each column characterizes a time slot of 15 minutes. The first row of the matrix $M_{N_{OD}}$ ($i = 1$ in Equation 3.13) reports the OD flows at every finest time step of the day (every 15 minutes). The following rows ($i \in [2, 6]$) refer to flow measured over largest interval periods, corresponding to the row temporal resolution. Each cell refers to the OD flow measured over the period matching the row resolution that include the considered 15-minutes time slot. For instance $m_{2,3}$ corresponds to the total flow observed during the 30-minutes period that includes the third 15-minutes slot of the day ([0.30 am, 0.45 am]): therefore, it measures the flow over the time slot [0.30 am, 1 am]. This progressive aggregation results in row blocks that are increasingly large, where the flows observed are the same for two time slots included in the same greater resolution period. Figure 3.13 provides an illustration of this matrix. For the sake of readability, the illustration focus on only eight consecutive 15-minutes time slots.

Based on this model, we also define $M^{OD,p} = (m_{i,j}^{OD,p}) \in \mathbb{R}^{7 \times 360}$ to characterize path flow dynamics along $p$ at different time slots and for different resolutions:

$$m_{i,j}^{OD,p} = \begin{cases} N_j^{OD,p} & \text{if } i = 1 \\ \sum_{j \in ]n \cdot 2^{i-1}, (n+1) \cdot 2^{i-1}]} N_j^{OD,p} & \text{if } i \in [2, 6] \\ \sum_{j \in [1,360]} N_j^{OD,p} & \text{if } i = 7 \end{cases} \tag{3.14}$$

For the sake of readability, we further let $m_{i,j}^{OD,p} = m_{i,j}^{p}$

Based on the definition of these two matrices, we propose an iterative algorithm (Algorithm 2) for estimating, for each origin destination couple $OD$, and for each 15-minutes slot $t$, the path flow distribution $P_t^{OD} = \{\alpha_{p,t}^{OD}, \forall p \in \mathscr{P}^{OD}\}$, estimated from the finest resolution ensuring the required statistical representativity.

**Algorithm 2** Adaptative path flow distribution estimation algorithm

**Require:** $t_{OD} \geq 0$ ▷ Define statistical representativity threshold
**Ensure:**
  $T \leftarrow 360$ ▷ # of quarter of hours in a day
  **for** $t \leftarrow 1$ to $T$ **do** ▷ For every time step
    $r \leftarrow 1$
    **while** $m_{r,t}^{OD} < t_{OD}$ **do** ▷ Find the finest resolution allowing a sufficient statistical representatitivity
      $r \leftarrow r + 1$ ▷ Decrease temporal resolution
    **end while** ▷ Either an intermediate resolution was found or the 24-hour scale is selected
    **for** $p \in P_{OD}$ **do**
      $\alpha_{p,t}^{OD} \leftarrow \frac{m_{r,t}^{p}}{m_{r,t}^{OD}}$ ▷ Compute path flow distribution
    **end for**
  **end for**

Basically, for each 15-minutes time slot, the corresponding path flow distribution is calculated according to the finest temporal resolution that ensures an OD flow above a pre-defined threshold. In Figure 3.13, the gray highlighting illustrates how, for each time step, a different temporal horizon is considered for estimating the path flow distribution, when fixing the required frequency threshold to 20. For instance, between 8:00 and 8:15, the frequency of the considered OD is below 20 (11). Therefore, a larger resolution is needed for estimating the path flow distribution at this time step. Considering a 30-minute horizon allows to reach the frequency threshold, since 44 travels are observed on the OD between 8:00 and 8:30. Therefore, the path flow distribution at 8:00 will be estimated with this 30-minute resolution. Instead, at 8:15, 33 trips are observed in a quarter of hour. This is higher than the frequency required to consider a temporal resolution suitable for path flow distribution estimation. Therefore, the path flow distribution at 8:15 will be estimated on the basis of a 15-minute time horizon.

Figure 3.14 further illustrates this temporally adaptive approach. It displays, for 25 different ODs (rows), the temporal resolution based on which the dynamic path flow distribution is going to be estimated at each 15-minutes time slot of the two-months historical period (January and February 2020). Figure 3.15 illustrates the distribution of selected resolutions when processing users of a single (non-historical) day of data (March 2$^{\text{nd}}$, 2020).

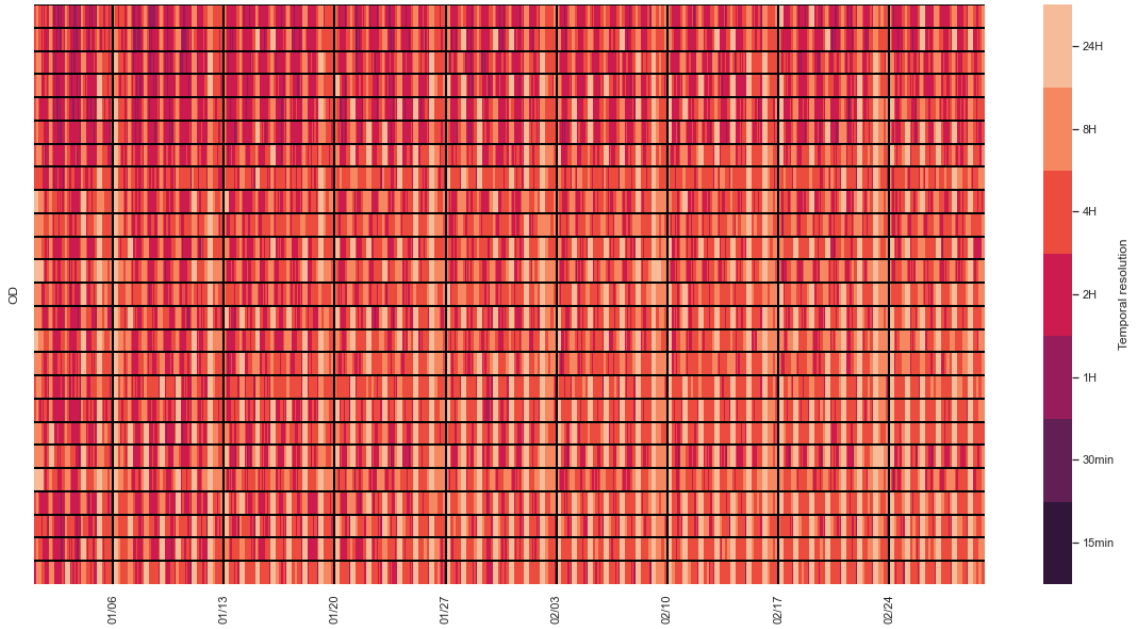



Figure 3.14: Resolution matrix for 25 different origin destination couples over an historical period of two months.

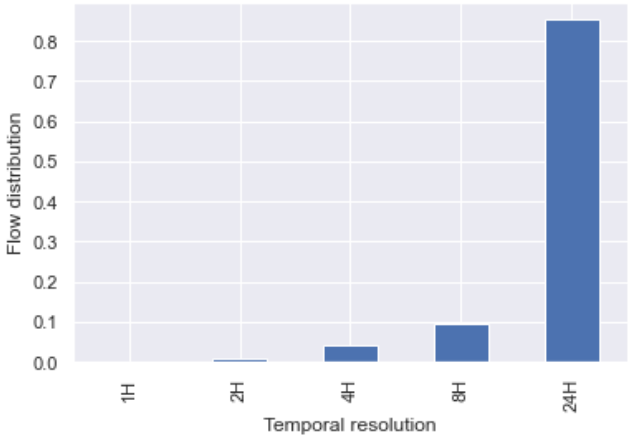



Figure 3.15: Flow resolution

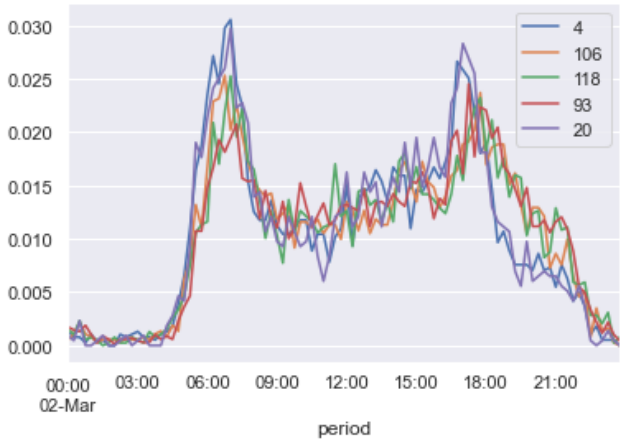



Figure 3.16: Different regional flow patterns for 5 regions

### 3.4.5 Path assignment and scaling

In Section 3.4.3, we proposed a way to map-match individuals at a regional level. Section 3.4.4 built on that work to estimate path flow distributions. To complement that work, this section develops on how these distributions can support the map matching the trips without pass-by point information, and the estimation of overall regional path flows.

Let $\mathscr{T}_d^{assign} = \mathscr{T}_d \backslash \mathscr{T}_d^{*\prime}$. Let $\mathscr{T}_d^{assign,OD}$ be the subset of $\mathscr{T}_d^{assign}$ made of trips traveling between $O$ and $D$. Let further $\mathscr{T}_{d,t}^{assign,OD}$ be the set of trips from $\mathscr{T}_d^{assign,OD}$ starting during the 15-minutes time slot $t$. We propose to assign each trip $\tau \in \mathscr{T}_{d,t}^{assign,OD}$ to a path $p$ by randomly sampling $p$ among $\mathscr{P}^{OD}$ based on the discrete probability distribution $P_t^{OD} = \{\alpha_{p,t}^{OD}, \forall p \in \mathscr{P}^{OD}\}$.

Applying this assignment step at each time step $t$ and for each origin destination couple $OD$ results in a new trip set $\mathscr{T}_d^{assigned}$. Merging this new set with the previously map-matched trips $\mathscr{T}_d^{*\prime}$ results in the overall map-matching of the trips of the daily regular users $U_d^r$.

In a final step, each trip is up-scaled (or down-scaled) according to the scaling factors that were estimated in Chapter 1. This final expansion allows to estimate overall path and regional flows. Figure 3.17 provides an illustration of the temporal variation of the path flow distribution for a given origin destination couple between three different paths. In Figure 3.16 instead, we compare the daily evolution the regional flow for different regions of the network, displaying different trends over morning and afternoon peaks.

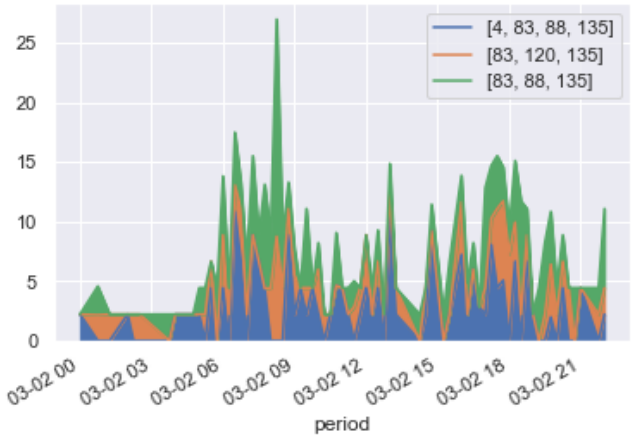



Figure 3.17: Path flow distribution for an origin destination couple

## 3.5 Validation perspectives and discussion

This chapter addresses the issue of estimating global regional traffic flows in the context of using temporally and spatially sparse data such as cell phone data. It proposes two enrichment methods with different scales of intervention. The first one aims at limiting the sparse character of the activity chains extracted from the telecommunication data. The second one aims at providing a complete identification of the routes taken by the users.

Both approaches raise the question of method validation. This issue is crucial to ensure that the estimated mobility is accurate. At the same time, it represents a real issue when using cell phone data. In this chapter, this validation step is limited to the comparison of origin-destination matrices. This step is all the more important as it is almost the only possible point between the interpretation of CDR data and the reality of mobility. We found that the measured origin-destination matrices are globally consistent with the field data, although some bias was observed. Future analyses should allow us to interpret and correct them. A similar validation of the path flows seems much more difficult to obtain. It would require additional data, such as GPS data to validate the path flow distribution, or count data to validate the regional flow measurement. At this stage, this validation is a significant challenge considering that these data are unavailable on the studied area.

However, upstream, a methodological validation of the developed methods could guarantee the robustness of the methods and thus increase the confidence in the observed results. This would be based, for example, on the selection of sufficiently rich individuals so that their mobility is artificially downsampled and then reconstructed and compared to their original mobility. The same approach could be implemented on GPS data that would provide detailed and reliable mobility information. This approach would be particularly relevant to evaluate the method of enrichment of activity chains, but also of the route taken by users. This will be the subject of future research.

## 3.6 Conclusion

This chapter targets two challenges raised by the mobility reconstruction from CDR data, related to the temporal sparsity of the data.

First, we note the sensitivity of mobility analysis to the amount of communication data. This raises a number of issues. At the individual level, it raises the question of detecting complete activity chains, and thus identifying movements. At the sample level, it raises the problem of correctly characterizing mobility, and in particular of estimating unbiased

origin-destination matrices. We address this problem by proposing a heuristic for enriching activity chains, based on the identification of mobility routines via a clustering method, and on the construction of individual mobility routines. It results in the enrichment of the activity chains of 11% of the daily regular population, with a satisfactory increase in flows during the day. However, we note an excessive enrichment of mobility at night and on weekends, which we explain in particular by the implementation of completeness thresholds unsuited to certain mobility profiles, and mobility routines constructed without distinction of the context and in particular the day of the week.

Second, we pursue the issue of the scarcity of cell phone data in the identification of individual trajectories, and thus in the reconstruction of mobility flows. We address this problem by developing an efficient map-matching technique, based on an exogenous identification of the prevailing paths enriched CDR-derived most significant paths alternatives. This enhanced prevailing path dataset provides a territory- and network-specific insight into the urban mobility. It allows the cost-efficient map-matching of users with little trajectory information onto the road network. The map-matched trips are further integrated into a path flow distribution estimation method based on flexible and variable time resolutions. This process aims at ensuring a minimal statistical reliability of path flow distribution, and hence of the path assignment later applied to trips lacking trajectory data. This work flow results in the evaluation of extended population path and regional flow estimation.

Because part of the methodology implemented in this chapter relies specifically on the identification of historical routines, this chapter focuses on a subcategory of the population identified as regular. In the following chapter, we will discuss estimating the traffic volume of the remaining irregular populations, through the scope of a more spatially aggregated approach.

# Chapter 4

# Irregular Mobility Reconstruction

## 4.1 Introduction

Recurrent activity patterns characterize a significant share of users. However, the traveler population also include mobility outliers. For instance, visitors (especially tourists) and some locals (taxi or delivery drivers) are likely to have a limited individual regularity. In the meantime, these users can be even more mobile than other regular users and contribute all the more to pollutant emissions. Some mobility studies exclude them from the analysis; such a choice is necessarily a limitation when estimating emissions. However, their irregularity prevents using historical data for mobility routine construction, automatic destination detection such as workplace, and mobility completion. Besides mobility completion, irregular mobility also raises the question of users' scalability, as these users' trips are not, by definition, representative of the rest of the erratic user's mobility. Therefore, completing the trips of erratic users (as we did for regular users in Chapter 3) is not only highly challenging but also unjustified given our objective.

Instead, in this chapter, we propose a way to include these users in our analysis without applying the traditional user-centered approach. Tourists, taxi drivers, or deliverers might have very changing itineraries from one day to another and from one individual to the other. However, they participate in more stable global mobility patterns, depending on the (economic or touristic) attractiveness of areas served. Therefore, we propose to directly compute an estimation of the traffic volumes.

This perspective raises different challenges.

1. First, it requires limiting the bias of the incompleteness of the data, while considering that an individual reconstruction of the data is irrelevant considering their irregularity. Therefore, it is a question of identifying representative users, which is a challenge.

2. Second, it requires reconsidering the spatial dimension of the mobility analysis. In Chapter 3, an analysis at the regional path level was relevant because of the likely representativeness of regular individuals within their own population. For non-regular individuals, the representativeness of mobility behaviors, at least at a fine scale such as that of regional paths, is questionable. It therefore seems appropriate to move to a less detailed scale of analysis, from which we can draw more legitimate conclusions.

3. With this change of scale, the question of the performance of the estimation of the calculation lengths arises again. Our method pays particular attention to this point.

The objective of this chapter is therefore to propose a complementary distance estimation method, adapted to a non-regular population. In this direction, Section 4.2 establishes the methodology. Section 4.3 presents the results, that integrate a sensibility analysis, a

validation study, and an exploration of the mobility patterns derived with the method. Section 4.4 concludes on the contributions of this chapter.

## 4.2 Methodology

### 4.2.1 Method outline

The methodological approach we propose relies on the following steps. A paragraph in this section is dedicated to each.

1. Selecting a subsample of users with large completeness levels for ensuring a robust individual distance estimation (Section 4.2.2);

2. Estimating the trip distances at the city scale based on an hybrid distance estimation method (Section 4.2.3);

3. Up-scaling the distances in order to represent the total irregular population $\mathscr{U}$ (Section 4.2.4).

In Section 4.2.5, we provide an approach to validate this method.

### 4.2.2 User Selection

In this thesis, we have repeatedly discussed the sensitivity of mobility analyses to the completeness of telephony data, and how it was especially a challenge when addressing the estimation of travel distances. Given the lack of regularity of the individuals this chapter focuses on, it is out of the question to try to enrich their mobility on the basis of an historical analysis. Yet, measuring the distance traveled by these users without such an enrichment step would result in a biased estimation of the distances traveled, lowered by the least active users in the population.

While the previous approach focused on an individual enrichment of the population, this chapter relies instead on a more collective approach. It is assumed that the distance travel by users is independent of their communication rates, although the literature has demonstrated a correlation between communication and the traveled distances [Couronne et al., 2013]. This point is further discussed in the conclusion of this chapter. Based on this assumption, we look into identifying a minimum sequential completeness threshold $\rho_c^{min}$ (cf. Section 3.3.1, Definition 8) above which the travel distances are well estimated. Let $\mathscr{U}_d^{irr}$ be the set of irregular users observed on day $d$. Let $\mathscr{U}_d^{irr,a}$ be the subset of irregular users observed on day $d$ that have a completeness above $\rho_c^{min}$:

$$\mathscr{U}_d^{irr,a} = \{u \in \mathscr{U}_d^{irr} |\ u.\rho_c > \rho_c^{min}\} \tag{4.1}$$

On the contrary let $\mathscr{U}_d^{irr,b}$ be the set irregular users observed on day $d$ that display an insufficient data completeness. We propose to extract the sub-population $\mathscr{U}_d^{irr,a}$. The further mobility analyses will be conducted on this users subset, that will be considered representative of the least active users $\mathscr{U}_d^{irr,b}$. Then the conclusion derived will be expanded to the global population $\mathscr{U}_d^{irr}$.

The selection of the activity threshold on the basis of which to separate the data requires a specific sensitivity analysis. This is presented in Section 4.3.3.

### 4.2.3 Distance calculation

The inclusion of non-regular users raises the question of how representative their mobility patterns are of the population they represent, considering a likely higher heterogeneity between non-regular users than between regular users. While it can be assumed that the distances they travel are representative, the question arises as to the origin-destinations and paths they travel along. These considerations suggest that a finely spatialized approach is not the most appropriate for treating these users. We therefore propose to work on a more global scale, that of the city of Cali ($Z_0$) and its surrounding area ($Z_1$).

The challenge is to estimate the speeds traveled by the users in $\mathscr{U}_d^{irr,a}$ at these scales based on their communication events at the base station scale. A simple way and fast way is to measure the Euclidean distance $d_E$ between the consecutive base station they visit:

$$d_E(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \tag{4.2}$$

where $p = (x_p, y_p)$ and $q = (x_q, y_q)$ characterize two geographic positions on the network. However, this approach is restrictive as it does not consider the road network structure neither and the resulting detours. Although it has been demonstrated that the shortest paths do not necessarily correspond to the paths taken by travelers [Yang et al., 2018], they still provide more accurate estimation of the distance traveled by users. Yet, as it was already discussed in Chapter 2, the computation of shortest paths can be costly. This is why we propose an alternative approach to estimate the shortest path distance $d_{SP}$ without actually running new shortest paths calculations.

We propose a hybrid approach based on two different methods depending on the scale considered: If the Euclidean distance between the two points considered is below a threshold $d_{min}$ In practice, the distance $d_H$ can be expressed as follows:

$$\begin{aligned} d_H &: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R} \\ (p,q) &\mapsto \begin{cases} d_{H,1}(p,q) & \text{if } d_E(p,q) \geq d_{min} \\ d_{H,2}(p,q) & \text{if } d_E(p,q) < d_{min} \end{cases} \end{aligned} \tag{4.3}$$

where $d_min$ is set to 1 kilometer. Below, we explicit the expressions of $d_{H,1}(p,q)$ and $d_{H,2}(p,q)$ and justify these choices. Figure 4.1 provides an illustration of this hybrid distance calculation.

#### 4.2.3.1 Case $d_E(p,q) < d_{min}$: a detour-ratio based function

The calculation of distances in the case of large distances is based on the concept of detour ratio, first introduced by Yang et al. [2018]. The initial intention of this notion is to relate the real distance $d$ traveled by humans with reference distances, such as the Euclidean distance $d_R = d_E$ [Yang et al., 2018, Paipuri et al., 2020], or the shortest path distance $d_R = d_{SP}$ [Paipuri et al., 2020]. Then, the detour ratio is defined as the ratio of the real distance $d$ with the reference distance, and characterizes the extra distance traveled with an observed route compared to the considered baseline.

Here, we propose to divert this notion, by defining a detour ratio $\rho_D$ as the ratio of the shortest path distance $d_{SP}$ with the Euclidean $d_E$:

$$\rho_D = \frac{d_{SP}}{d_E} \tag{4.4}$$

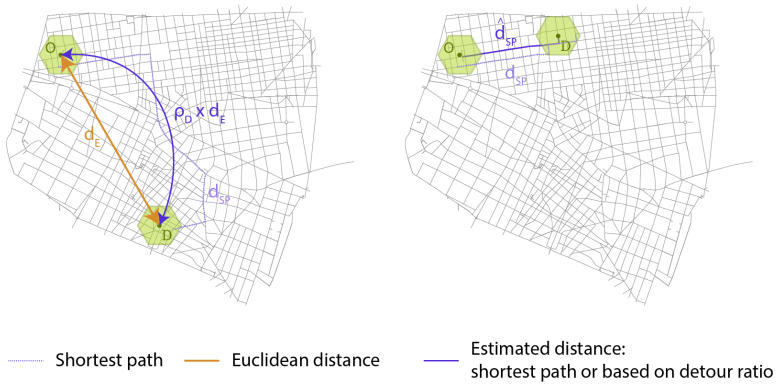



Figure 4.1: Illustration of the hybrid approach for estimating the distances traveled by an individual between two base stations. Left: For long distances, we resort to an approximation of $d_{SP}$ through the use of a detour ratio function and the Euclidean distances between the base stations. Right: For short distances, we proceed to a calculation of the shortest paths, allowed by the analysis of limited geographical areas.

Expressing $\rho_D$ as a function of $d_E$, and calibrating the relation, it provides a simple way to estimating the shortest path distance associated with a given couple of geographic positions $(p, q)$ from the Euclidean distance between them:

$$d_{H,1} = \hat{d}_{SP}(p,q) = \rho_D(d_E(p,q)) \cdot d_E(p,q) \tag{4.5}$$

To calibrate this relation, we suggest to exploit the set of shortest paths generated for regional paths characterization in Chapter 2. The detail of this calibration are presented in Section 4.3.1.

#### 4.2.3.2 Case $d_E(p,q) \geq d_{min}$: a simple shortest path calculation

The method selected for calculating $d_{H,1}$ provides background for a light-weighted, easy-to-implement estimation of shortest-path distances between a set of points. However, one limit of the method is that the detour ratio presents a significant variability for short Euclidean distance, where it should not be trusted. Therefore, we complement the detour ratio approach with a more sensible approach for small Euclidean distances. For such limited distances, we simply resort to a shortest path distance calculation, facilitated by the analysis of a limited spatial perimeter.

As a consequence, Equation 4.3 can be re-written as:

$$\begin{aligned} d_H &: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R} \\ (p,q) &\mapsto \begin{cases} \rho_D(d_E(od)) \cdot d_E(od) & \text{if} d_E(od) \geq d_{min} \\ d_{SP}(od) & \text{if } d_E(od) < d_{min} \end{cases} \end{aligned} \tag{4.6}$$

The calibration of the deviation ratio function (Section 4.3.1) gives insight into a reasonable value for $d_min$. For two points separated by an Euclidean distance above $d_min$, the detour ratio is trusted and used for the distances estimations. Below this threshold instead, we process to a simple shortest path distance computing, with temporary spatial network restrictions to fasten up the running.

The function $d_H$ we defined provides a way to efficiently estimate the distances traveled by users on the network. Due to the spatial inaccuracies of the data, we apply this metric to measure the distances between consecutive activities of the users' activity chains. Pass-by points are therefore not included in the distance estimation. Their inclusion in the approach would have the consequence, especially in the case of rich traces, of introducing an accumulation of estimation error. In future improvements of this method, it could nevertheless be relevant to downsample the traces, in order to select only the pass-by points providing additional information on the distance traveled.

Therefore, estimating the daily distance $d_u$ traveled by a user $u$, of spatial chain $c = (l_1, ..., l_n)$ can be formulated as the sum of the hybrid distances between its consecutive activities:

$$d_u = \sum_{i}^{n-1} d_H(l_i, l_{i+1}) \tag{4.7}$$

At this stage, in order to estimate the distance traveled within the urban area $Z_0$, which is the targeted area for emission calculation, we need to allocate an estimation of the share of the trip length between $Z_0$ and $Z_1$. To do so, we consider the geometrical line joining the origin to the destination, and measure which proportion $\alpha$ of this line is included in $Z_0$. The corresponding share of the distance is allocated to $Z_0$, the remaining share is allocated to $Z_1$. In future works, we will reflect on how to distribute more precisely the trip lengths over the different regions of the network, based on the network density and observations from the regular population.

### 4.2.4 Scaling

Once the daily travel distances of individuals in $\mathscr{U}_d^{irr,a}$ have been estimated, we perform a double scaling to extrapolate the conclusions the population it represents. First, the individual distances are up-scaled according to their own individual weighting, as estimated in Chapter 2:

$$TTD^{irr,a} = \sum_{u \in \mathscr{U}_d^{irr,a}} TTD_u * s_u \tag{4.8}$$

Then, the resulting total distance is further up-scaled in order to represent as well the population that was filtered because of two little data completeness.

$$TTD^{irr} = TTD^{irr,a} \cdot \frac{\sum_{u \in \mathscr{U}_d^{irr}} s_u}{\sum_{u \in \mathscr{U}_d^{irr,a}} s_u} \tag{4.9}$$

This second expansion can be discussed, in view of the biases existing between the communication activity and the mobility of individuals. At this stage, the hypothesis of a good representativeness seems to us to be satisfactory in view of the objectives. It could be reconsidered in future improvements of the method.

### 4.2.5 Validation

Using a detour ratio function to estimate shortest-path distances necessarily results in errors in the distances estimations. The local variability of the coverage of cell phone antennas may also contribute to errors. This section proposes a simple validation framework to evaluate these error levels.

It relies on generating synthetic trips at the scale of the road network, and comparing the distances evaluated with our method against the real network distances.

The approach involves the following steps:

1. Random sampling of origin-destination pairs on the road network, and the computation of shortest paths between these origins and destinations;
2. Computing the shortest path for each origin-destination pair;
3. Estimating the travel distance with $d_H$;
4. Comparing the results with the reference distances at the network level.

In order to compare the performance of $d_H$ with a simple Euclidean-distance-based estimation, we repeat the experiment with $d_E$. The result of this analysis are presented in Section 4.2.5.

Here, it is important to emphasize that the measured errors correspond to deviations between an estimate of the shortest path distances, and the shortest path distances themselves. It is therefore not a question of validating that the estimated lengths correspond to the path lengths of the users of CDR data on the Cali network, which we try to approximate throughout this thesis by analyses based on the assumption that the shortest path is a good approximation of the length actually traveled.

## 4.3 Results and applications

### 4.3.1 Detour calibration

The set of routes generated by the calculation of regional lengths is used for calibrating the detour ratio function in Cali. In a first step, we compute the total trip distance from the shortest route, as well as the Euclidean distance from the origin and destination nodes positions. It allows associating each route with a detour ratio. Averaging the detour ratio distribution by step of 500 meters returns the blue distribution in Figure 4.2. We found this curve can be fitted by :

$$\rho_D = 1.132 + \frac{0.872}{d_E + 0.548} \tag{4.10}$$

with $R^2 = 0.97$.

Interestingly, the average detour ratio $d_r$ is characterized by slight oscillations. These oscillations are a consequence of the bias we introduced by using the border-to-border shortest paths sample. As the regional partitioning involved a minimal surface criteria, the borders from one zone to another are separated by quasi-regular distances. It results in a periodical over-representation of some distances compared to others, which affects the distribution of the detour ratio. The distribution of the Euclidean distance between the shortest paths set is displayed in the background of Figure 4.2. It shows that the peak in length frequency coincides with the decrease in standard deviation. Although a more uniform sampling would likely return a smoother curve, this would require shortest path calculations to be performed again. At this stage, this would most certainly have little impact on the calculation of distances, and we leave this perspective for future work.

It is also worth mentioning that we rely on itineraries taking place in various contexts: travel in urban areas, commuting between distant centralities, travel in mountainous or more rugged areas. In principle, one could therefore calibrate detour functions according to the origin-destination pairs considered, and obtain more specific results and better fits. Once again, for the moment, we will be satisfied with a single detour function, which fits well to the data set.

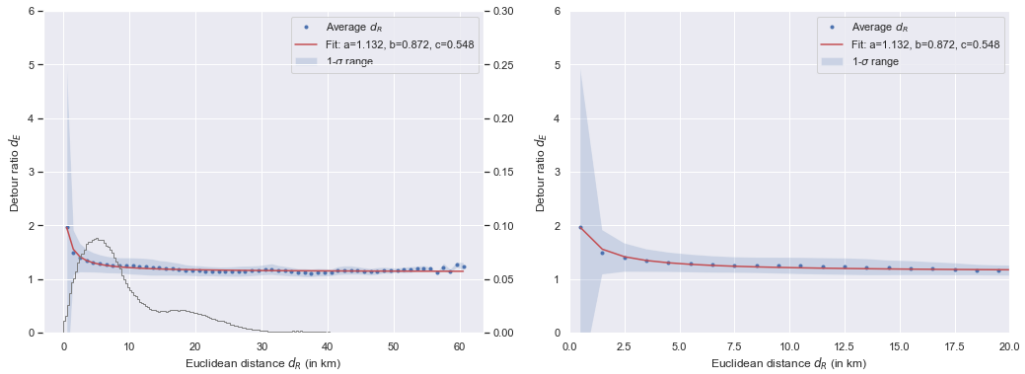



Figure 4.2: Detour ratio function calibration: global and zoomed in plots

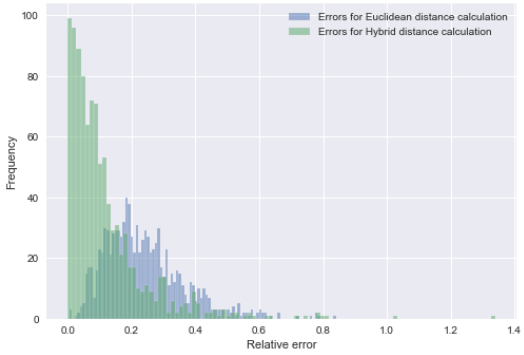



Figure 4.3: Relative error distribution with hybrid distance and Euclidean distance

### 4.3.2 Validation

This sections presents the results of the implementation of the evaluation process of the method described in Section 4.2.5. The objective is twofold. Firstly, it is a question of evaluating the estimation errors introduced by our method, in comparison with the lengths actually traveled. Secondly, it is also a question of identifying the gain compared to a simple calculation of distances based on Euclidean distances.

The results presented in this section are based on the generation of 2000 artificial trips.

Figure 4.3 displays the distribution of the errors to real shortest path for distances estimates based on $d_H$ and $d_E$. It is clear that errors are much more limited with the hybrid distance approach.

Figure 4.4 further compares the evolution of the average of relative errors with the Euclidean distance between origin and destination points. On average, the deviation from the $d_{SP}$ distance is twice as small with the $d_H$ distance than with the Euclidean distance, and remains contained around 10%.

### 4.3.3 Sensibility analysis

In this section, we discuss the sensibility of the distance estimate with the daily number of events observed. For an sample of 30,000 users, we compute the daily number of events, and relate it to measured traveled distances. We split the number of events in even intervals and compute, for each, the average daily distance traveled by the corresponding users. They are displayed in Figure 4.5, along with the cumulative distribution of the daily number of events. Up to a completeness of 0.6, we observe a linear growth of the average daily distance traveled with the daily number of events. After this completeness threshold, we observe a stabilization of the average travel distance at a level of approximately 60 kilometers. This

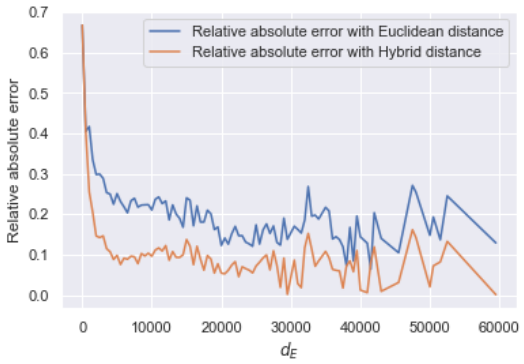



Figure 4.4: Average relative error evolution with Euclidean distance $d_E$

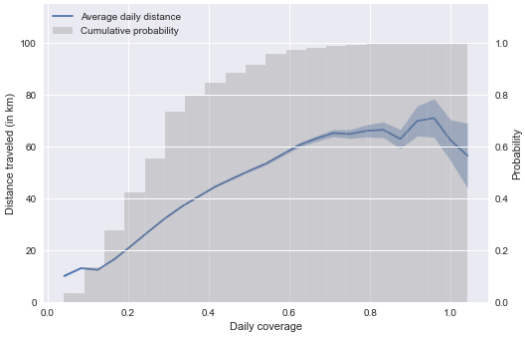



Figure 4.5: Irregular population: normed category's total traveled distance

stabilization seems to confirm that beyond a certain completeness threshold, the estimated distances are well represented: more data does not mean more distance traveled. This 0.6 value is on a daily basis to set the completeness threshold $\rho_c^{min}$ used to separate $\mathscr{U}_d^{irr,a}$ and $\mathscr{U}_d^{irr,b}$.

### 4.3.4 Application: historical analysis

In this section, we explore the results of the application of the detour ratio approach for estimated distances traveled by users. We apply the method to the historical period, which has sufficient temporal depth to evaluate the relevance of the method. We first evaluate the total travel distances for the irregular population, then further investigate the distances traveled within the overall population for each different user category.

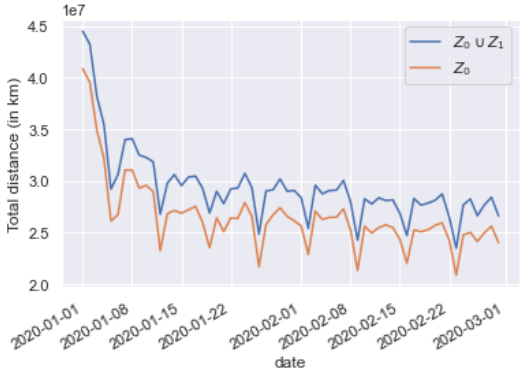



Figure 4.6: Irregular users: total travel distance

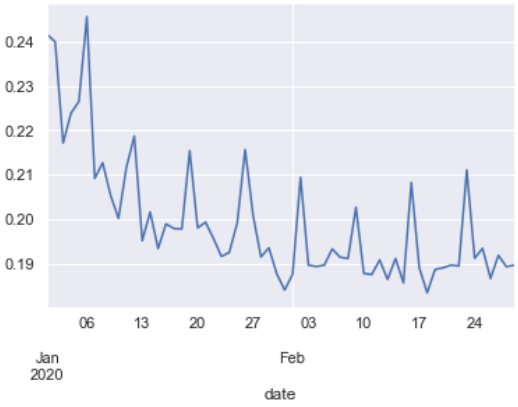


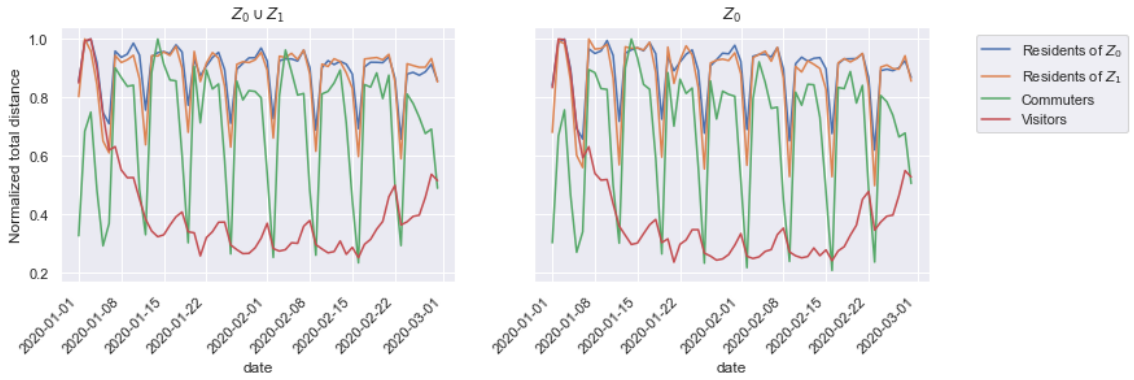

Figure 4.7: Irregular users: weight in the overall population TTD


Figure 4.8: Overall population: total travel distance per category

Figures 4.6 the total travel distances estimated for the irregular population $\mathscr{U}^{irr}$. The blue line represents the trend in the overall region $Z_0 \cup Z_1$, while the orange line represents the distance traveled within the city of Cali only ($Z_0$). It appears that the city of Cali represents a very significant share of the distances traveled in the agglomeration. There are clear weekly seasonalities, with distance drops on Sundays. coupled with a notable increase in the distance travelled at the beginning of the study period, which corresponds to the New Year celebrations and vacations. This increase can be explained by two factors: first, a probable influx of individuals considered irregular, the visitors; second, a possible increase in the average distance travelled by users during this holiday period. The amount of kilometers traveled is significant. When compared to the total distance traveled, we estimate that the irregular users contributes to 19 to 25 percent of the mobility volume. Figure 4.7 illustrates this trend. Interestingly, it shows how the irregular users have an increased weight in the global mobility on a weekly basis as well. This weight rises every Sunday, which means that even though irregular users are relatively less mobile on Sundays, their mobility drop is weaker than for the regular users. Together, these plots show that irregular users contribute to an important share of the total traveled distance, and therefore that they should not be neglected.

In Figure 4.8 is displayed the total traveled distances of the overall population, separated by presence profile (residents of $Z_0$ and of $Z_1$, external commuters, and visitors). From the outset, we observe the significant difference of magnitude between the total traveled distance by residents, and other users categories. This observation can legitimately question the user categorization approach. However, a deeper investigation showed that commuters and visitors cumulated close to 300,000 kilometers on a daily basis on the over-


(a)




(b)

Figure 4.9: Normalized total travel distances of the overall population: (a) with completeness-based user selection (standard method), (b) without completeness-based user selection

all metropolitan area of Cali, which is considerable. Also, we noted a significant impact of the completeness filter on commuters, drastically reducing the sample size. It is likely that the results for this part of the population are less reliable. This limitation of the completeness filter has already been discussed in the previous chapter, and should be addressed in future improvements of the method. For the overall population, the visualization of the total travel distance also plays a weekly regularity.

For a better identification of the seasonal patterns and how they are related from one user category to the other, Figure 4.9 displays, for the overall population, the total travel distance normalized by the maximal daily distance of the category. This allows us to compare patterns on a similar scale. We propose two versions of this graph, the first one in the standard case where the population is reduced to the users with a completeness higher than 0.6 (Figure 4.9a), the second one based on all the users, whatever their completeness level (Figure 4.9b). The purpose of this distinction is to demonstrate that although commuters are poorly represented in the population reduced to the most complete users, they actually have mobility patterns as regular as the other categories of the population. In Figure 4.9b, we clearly observe the temporal shift between the profiles of the mostly regular users (residents and commuters) and the profiles of the visitors. In the version reduced to a subsample of the population (Figure 4.9b), the temporal regularities appear less clearly. Due to smaller sample sizes and less robust expansion processes, the results are more sensitive to individual variations. The integration of larger amounts of data should limit the noise observed on the curves.

Figures 4.10 displays, for the overall user population, the average distance traveled by


(a)


(b)

Figure 4.10: Average travel distances of the overall population: (a) with completeness-based user selection (standard method), (b) without completeness-based user selection

users for each considered mobility profile, both in the overall studied perimeter ($Z_0 \cup Z_1$) and in the inner city ($Z_0$), and the 95%-confidence interval. We can make several observations. First, the considered scale influences the ranking of user categories. Based on a restricted sample analysis (Figure 4.10a), residents from $Z_0$ and $Z_1$ are observed to travel equivalent distances (60 kilometers on average). At a smaller scale, the two resident categories are more separated, with longer distances traveled by residents of $Z_0$, which suggests that the residents from Cali's surroundings tend to travel more in the surrounding area as well. Interestingly, Figure 4.10b provide a different point of view for the ranking of the distances traveled in $Z_0 \cup Z_1$. When not restricting the analysis to the most complete activity chains, residents from $Z_1$ are estimated to travel more kilometers on average than residents from $Z_0$. This difference shows the sensitivity of the analysis to filters on communication, and suggests different communication properties between users of $Z_0$ and $Z_1$. In $Z_0$ instead, we observe that the same ranking than previously, completed with a clearer positionning of the visitors and commuters. Surprisingly, commuters do not exhibit travel distance behaviors closer to the residents of $Z_1$, to which they could be related.

Generally speaking, and for both regions ($Z_0 \cup Z_1$ and $Z_0$): around 40 to 60 kilometers per day if we consider the estimates based on a restricted sample. These values are above our expectations. However, they are drawn up by the most mobile users, while the median is actually lower. In future works, we would like to better investigate the properties of each considered sub-population in order to better explain such specific foundings.

## 4.4 Conclusion

In this section, we propose a method for estimating travel distances for individuals whose irregularity does not allow for individual mobility reconstruction. The method we propose thus uses a collective reconstruction, in the sense that a sub-sample of individuals supports the estimation of the distances traveled by the whole group. The sample in question is selected according to a completeness threshold, set following a sensitivity analysis. The objective is to select a threshold beyond which it is identified that the estimated distances are relatively stable. Below this threshold, on the other hand, there remains a sensitivity to the users' communication activity that is detrimental.

For selected users, we propose a hybrid distance estimation method. Maintaining our assumption that the shortest path distance is a good approximation of the distance actually traveled by the users, we propose a method to estimate this distance without resorting again to shortest path calculations. This is based on the notion of detour ratio, which measures an extra distance from a reference. We show that this method generates limited errors in the estimation of the shortest path distance, and that it produces much finer results than a simple application of the Euclidean distance.

Considering the likely heterogeneity of irregular users' behaviors, we choose to estimate distances at a macroscopic scale. This choice guarantees a better scalability of the results. This difference in scale with the variables estimated in the previous chapter is not problematic in itself. Nevertheless, in future work, we can try to disaggregate this information to finer spatial and temporal levels. The results of Chapter 3 could be used as a model to weight and redistribute total distances traveled over the day and over the territory.

Since the literature has shown a correlation between travel distances and user communication levels, the assumption that the mobility of users that communicate the most is representative of users that communicate less will have to be reconsidered at some point and to some extent. A specific study could, for example, make it possible to call for appropriate corrective factors to avoid overestimating the distances traveled.

## Conclusion of Part II

This second part of the thesis focuses on estimating path flows and total distances traveled on the network. Starting from the observation that mobile phone data alone, which are biased, are not sufficient to accurately estimate the mobility of the population, enriching these data is necessary. To this end, we propose two parallel and concurrent approaches, each adapted to a different population category.

For a favorable part of the population, *i.e.*, individuals with a data history characterized by regular mobility, we propose to use a reconstruction of mobility based on historical knowledge. This reconstruction involves two levels of analysis. On the one hand, partial daily activity chains are enriched based on individual historical routines considered as reliable. On the other hand, the trips between activities are enriched using the prevailing paths knowledge. Depending on the level of information available for a trip, assigning the user to a path can either:

- Be exclusively based on individual data.
- Mix individual data with prevailing paths information.
- Rely on a collective approach based on path flow distribution estimation.

*In fine*, this approach results in a dynamic path flow estimation. Crossed with the average distances traveled per path, it allows computing the total distances traveled per region and time interval.

However, this first approach is only valid for a part of the population. A certain number of individuals cannot be targeted by this approach, either because their history is too sparse (visitors) or too erratic (non-regular users). The notion of a historical mobility routine no longer makes sense for these individuals. Therefore, we propose an alternative approach for estimating the distances traveled. The irregularity of individual behaviors suggests low representativeness of the trajectories taken by these individuals. For this category of the population, we propose to change the scale of analysis and to estimate the total distances traveled at the scale of the entire urban area and the day. To limit the impact of communication bias on the estimation of individual distances, we focus on a subgroup of this population characterized by high levels of communication, assumed to be representative of the rest of the target population. The individual distances traveled by these users use a hybrid metric based on the notion of detour ratio. Applying this method to the targeted sub-sample and expanding the results provides an estimate of the total distances traveled by this irregular portion of the population.

The significant contributions of this section are, therefore:

- Relating population classification methods with different and competing mobility reconstruction methods. This innovative articulation allows a systematic mobility reconstruction, whatever the user profile.
- Developing a heuristic for enriching individual activity chains based on identifying historical mobility routines.
- Implementing a regional map matching method. It is based on a preliminary knowledge of regional paths enriched by data-centered observation. It involves assigning trips to a route with a differentiated approach depending on the availability of trajectory data.
- Adapting and using the detour ratio concept to estimate irregular users' distances traveled on the network.

Some of the research directions discussed in this part of the thesis are the following:

- Further adapting the activity chains completion method to each mobility profile, especially by making the approach less dependent on high activity communication rates.
- For the irregular mobility reconstruction, conducting a specific study to identify the bias on distance introduced when focusing on most active users. Proposing a correction factor to limit its impact on the traveled distance results.
- Developing methods for re-distributing the traveled distance in space and time to refine the analysis resolution.

# Part III

# Prototyping a traffic-related emission calculation tools based on CDR data

## Introduction

The previous part of this thesis dealt with traffic volume estimation. This traffic variable alone is insufficient to estimate traffic-related air emissions. Indeed, it only allows the expansion of the emission factors, which depend on the average traffic speed. The objective of this part is, therefore, twofold. First, we intend to propose a method to target the speed estimation from CDR data, to support the estimation of emission factors. Then, we aim to present a general data processing and modeling framework that relates this traffic speed estimation method with the other processes presented previously in this thesis.

Chapter 5 targets the speed estimation issue. Considering the mobile phone data characteristics, and especially the temporal sparsity, it is a challenging one. The method we propose relies on data fusion to overpass the temporal data biases. This chapter is an updated version of the journal paper:

> Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, and Thamara Vieira da Rocha. Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths. *Transportation Research Part C: Emerging Technologies*, 129:103183, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103183. URL `https://www.sciencedirect.com/science/article/pii/S0968090X21001996`

The work presented in this paper was conducted prior to the reception of CDR data, which occurred 15 months into the thesis. Therefore, the method was designed to handle a generic data format, called user-activity dependent positioning (UADP) data. The method was tested on a different case study from Cali, over the city of Lyon, France, where it was applied onto artificially-generated trips from floating car data. Due to time constraints, the experiment could not be repeated on real data and the city of Cali. This study is an obvious perspective of prolongation of the work of this thesis.

Chapter 6 aims to overcome this difference in the territories analyzed, by proposing an integrated methodological framework that links together the unit method developed in this thesis. In this chapter, we also discuss the most significant research gaps remaining to fill for implementing and using this framework in practice.

---

Outline

# Chapter 5

# Speed dynamics estimation for generic user-activity dependent positioning data

## 5.1 Introduction

Over the last two decades, the digitalization of services and infrastructures has led to the emergence of a broad set of new information sources to characterize human mobility. In particular, GPS tracks from navigation systems and services have become prevalent [Castro et al., 2013, Lin and Hsu, 2014]. The exploitation of other sources, such as anonymous geolocalized social media logs (Twitter, Foursquare) and cell phone data, has also become increasingly popular [Chen et al., 2016]. The collected geolocated tracks may vary significantly in both the spatial and the temporal resolution depending on the technology used to generate the data (GPS, mobile telephony, wireless networks), the sensing device or service (on-board or mobile navigation systems, geolocation through social networks and location-based services, 2G, 3G 4G cellular networks), as well as the level of user activity [Asgari et al., 2013, Toch et al., 2018].

GPS tracks derived from vehicular and mobile navigation systems are usually quite accurate both in space and time. The navigation system generally acquires the user's position at a regular frequency, which usually ranges from a very high frequency (*e.g.*, every second) to lower sampling rates (in the order of the minute). Despite these possible variations and acquisition noise and errors, GPS navigation systems remain a key source to explore individual and aggregated mobility patterns and monitor traffic [Castro et al., 2013, Lin and Hsu, 2014]. However, the related data sets often suffer from limited penetration rates.

Another source of information on human mobility can be found in social networks and Location-Based (Networking) Services (LBS - LBNS). For instance, the Twitter social network allows users to share their geo-location with their tweets, while the LBNS Swarm (formally called Foursquare) offers its users to "check-in" in various venues and share this information with friends. GPS being the technology on which those networks and services rely, the spatial accuracy of the data generated using such services is mostly the same as that of navigation systems. However, the main difference with the latter lies in the data generation process. Instead of being automatic and regular, the availability of geolocated samples with social networks and location-based services depends on the user's communication and sharing behaviors. In particular, users with little posting and check-in activities will generate fewer location data, and their mobility becomes harder to estimate.

This is a characteristic that social networks and location-based services data share with several types of passive mobile phone data, such as Call Detail Records (CDR) and network

signaling data (which, in addition to calls and texting events, include network control ones such as handovers). These data are passively generated by mobile phone users while communicating and are collected and stored by communication data providers for billing or network management purposes. CDR data register each communication event (*i.e.*, a call, message, or data browsing event emitted or received by a cellular device) at the base station scale (*i.e.*, antenna), while handovers register each base station involved in a call. Thus, the less a user communicates, the fewer data will be generated. Barabási [2005], and later Candia et al. [2008], Gonzalez et al. [2008], Calabrese et al. [2011] and Chen et al. [2018], explored the existence of patterns in mobile communication behaviors and observed that the latter are bursty. While most of the users' communication events happen within short time intervals, some significant time gaps also exist between successive dense communication sequences. Interestingly, Gandica et al. [2017] demonstrated that message posting on Twitter presents similar temporal characteristics. Those results suggest that these user-activity-dependent positioning data (UADP data further on) may be more fitted to identify and analyze the static phases (often called *stays*) of users' routines than to characterize the trips in-between [Ranjana et al., 2012, Hoteit et al., 2017].

An extensive literature exists on the use of user-activity-dependent positioning data for mobility analysis (see Blondel et al. [2015], Naboulsi et al. [2016] on mobile phone data), but it mainly focuses on the characterization of mobility patterns rather than the analysis of dynamic traffic features. This literature is often based on methods to detect and process communication events that take place during periods of human immobility (*e.g.*, see Jiang et al. [2013], Toole et al. [2015]), which allow inferring origin and destination locations of trips, for instance. On the basis of such methods, the subjects covered by the literature vary from the exploration of mobility habits and characteristics (Jurdak et al. [2015] with Twitter data) and the development of realistic mobility choice models (Gonzalez et al. [2008] based on CDR data) to the construction of origin-destination matrices (see Osorio-Arjona and García-Palomares [2019] with Twitter data, Iqbal et al. [2014], Çolak et al. [2015], Alexander et al. [2015] with CDR data) and their use as a proxy for the traditionally costly transportation surveys. However, when it comes to describing the trips themselves, the irregularity of the communication behaviors and the individual data generation may result in little to no positional information during trips that is therefore much harder to exploit. Even if some data are collected during a trip and can help to identify the likely traveled routes, as shown in Jiang et al. [2013], this situation is far from being systematic and only concerns few positions. Due to this limitation, the studies related to dynamic mobility pattern characterization are less developed. In Toole et al. [2015], a CDR-based origin-destination matrix is estimated in a first step, then assigned onto the road network in a second step to estimate the traffic load. Handovers and Location Area Updates are used to estimate traffic speeds on highway segments [Bar-Gera, 2007, Ou et al., 2011], travel time [Janecek et al., 2015], or Macroscopic Fundamental Diagrams (MFD) [Derrmann et al., 2017]. However, handovers guarantee a minimum frequency of location updates during calls, which is not the case for other data sources such as traditional CDR or social media logs. Whether UADP data can still be used to derive dynamic traffic characteristics, such as speed, remains an open question.

Monitoring urban network traffic speed is crucial for many applications, including traffic control, route guidance, or emission calculations [Zhang et al., 2011]; and targeting speeds from irregular and low-frequency positioning data remains the most challenging application. In fact, the traditional bottom-up speed estimation methods from GPS floating vehicle tracks [Zheng et al., 2013, Shang et al., 2014], which rely on averaging individual speeds calculated at the road segment level, cannot be transposed to this kind of data. However, user-activity-dependent positioning data have significant advantages with respect to more conventional traffic data sources. They are usually accessible and massive. Mobile

phone data have very high penetration rates among the populations [Blondel et al., 2015, Algizawy et al., 2017, Bachir et al., 2017], which results in excellent spatial coverage in urban areas. Social network data are massive as well. They still offer lower penetration rates (because they correspond to more specific audiences and uses) than cell phone data, but their availability continues growing [Cisco, 2020], offering promising perspectives in more extensive use for mobility analysis. Traffic speed estimations based on GPS floating vehicle tracks often rely on complementary data sources (like surveys, loop detectors, or cameras) to implement spatial extrapolation processes and compensate for the low data coverage [Shang et al., 2014, Zhan et al., 2017], leading to costly overall processes. On the contrary, working with temporally sparse but massive data seems promising as it could offer cost-efficient and large-scale alternative methods. Given the massive amounts in which UADP data are available, and despite their temporal irregularity and sparsity, we aim to prove that they offer in an urban context a viable alternative to GPS floating car data to estimate the mean traffic speed dynamics at a zonal scale. By focusing on UADP data, we consider all massive mobility data related to the use of new technologies and whose temporal sampling frequency depends on users' communication behaviors and activities, and therefore inherently uncertain.

A key point of the method we propose is that it is based on the partition of the urban network into regions characterized by homogeneous traffic conditions. This partition defines a new spatial scale at which the individual trip data are up-scaled and analyzed. This aggregation process allows characterizing interrelated travelers, *i.e.*, travelers who simultaneously cross the same network areas, but is also more adapted to the possible raw spatial resolution of the data than the road segment scale. Thanks to this new scale, our method only requires a set of elementary trip features but no explicit characterization of individual local speeds. Those features are the observed departure and arrival time, and the regional path (as defined in Yildirimoglu and Geroliminis [2014], Batista et al. [2019]), *i.e.*, the succession of regions traveled by individuals between their origin and their destination regions.

We propose to fuse from the outset the travel time information of individuals traveling along the same regional paths on a periodic regular basis (*e.g.*, every 15 minutes), and conduct, for each of the considered period, a combined analysis of the average travel times estimates derived from this data fusion. Provided that a reliable estimation of the trip lengths at the city and regional scales is available from external offline sources, this analysis allows deducing a broad and consistent estimation of the regional average traffic speeds. One of the main challenges of applying this methodology is the correct estimation of average travel times, despite the temporal biases inherent to the use of user-activity-dependent positioning data. The method we propose relies on statistical considerations to addresses this challenge.

We apply the method to a set of artificially temporally-biased trips derived from a real GPS dataset of tracked vehicles traveling in the Great Lyon area, France. This approach allows using the original GPS dataset as a ground-truth reference for traffic speed, against which to assess the methodology and determine whether the simulated data are qualified for urban traffic speed estimation. Although the GPS dataset size is limited, literature works have shown that GPS floating car data was a particularly reliable source for estimating zonal traffic speed. Contrary to other traffic variables, the traffic speed estimation does not require scaling processes. Its estimation from vehicle probe data results in very satisfactory results despite low penetration rates [Nagle and Gayah, 2014, Leclercq et al., 2014]. In this research, by keeping the data downsampling process under control instead of directly using UADP data, we aim at better understanding how the jamming and the consequent progressive information loss could impact the quality of the results. By focusing on a synthetic data context that permits clear identification of the temporal bias, this study

aims to assess the robustness of the proposed methodology towards its application on non-synthetic UADP data.

This article is organized as follows. Section 5.2 exposes the principle of our approach and describes the proposed methodology. Section 5.3 presents our case study, as well as the exploited data. Section 5.4 focuses on the results we reached. Finally, Section 5.5 concludes with the limits and perspectives of this work.

## 5.2 Methodology

### 5.2.1 Problem statement

We focus on exploiting vehicle trips extracted from a generic UADP dataset (mobile phone data, LBNS data, or any similar mobility dataset) leveraging the literature stay detection methods. These methods define *stays* as locations (either a specific position or a cluster of positions close to each other) where users are observed for a minimum amount of time. These methods are therefore geared towards identifying static phases of individual mobility. We will consider that such methods are reliable and that the stays detected do indeed correspond to static phases. Nevertheless, this identification is dependent on the communication activity of the users. It implies that static phases may only be partially identified if the user is not active at the beginning or the end of their stay. Suppose we define a *trip* as the mobility phase between two consecutive stays. In that case, an important distinction must be made between the observed trip departure and arrival times and the exact (but unknown) ones, as the varying communication rates of users provide sparse information on their mobility. We use the following definitions, $i$ being an individual trip:

- The *observed departure time* is defined as the time when the last static event of the origin stay is observed. By definition, the observed departure time precedes the actual one. In this chapter, let $\epsilon_d^i$ be the positive bias between these two values (all mathematical notations in the article are listed in the notation table in Appendix C.1).
- Reciprocally, the *observed arrival time* is defined as the time when the first static event of the destination stay is observed. By definition, the observed arrival time follows the actual one. Let $\epsilon_a^i$ the positive bias between these two values.
- The *observed travel time* $T_{obs}^i$ is defined as the time elapsed between two consecutive stays, *i.e.*, between the observed departure and the observed arrival times. It is an overestimate of the actual travel time $T^i$.

Based on these definitions, we have:

$$\epsilon^i \;=\; T_{obs}^i \;-\; T^i \;=\; \epsilon_d^i + \epsilon_a^i \tag{5.1}$$

Intermediate trip positions can give additional information, considering that the departure time occurs between the observed departure time and the first mobile event. The reasoning is symmetrical for the arrival time. Therefore, the longer the delay between consecutive moving and static events, the more uncertain the departure and arrival times, the greater the risk of significant overestimation of the individual travel time. These individual biases are, by nature, very difficult to estimate at the trip level, and they affect the observations of the individual travel time themselves.

In this context, the problem we address is the following. Can we provide a method to correctly estimate traffic speed at least at an aggregated regional scale despite these unknown temporal individual biases?

### 5.2.2 Overview

The fundamental principle behind the speed estimation method we propose is that the overall sample size of the data can compensate for the low data quality at the individual trip level. The method relies on the fusion of individual trip information and statistical considerations to provide a reliable regional traffic speed estimation. It requires the implementation of the following steps.

1. Network partitioning and time resolution definitions;
2. Average de-biased travel time estimation through the periodic gathering of similar trips;
3. Speed calculation through the resolution of a linear system model;
4. Speed trends smoothing.

These steps constitute the generic skeleton of the methodological framework we propose. However, some of these steps will require adaptation to the specific properties of the input data and case study.

The network partitioning is the starting point of our methodology. It participates in the definition of the spatio-temporal resolution of the final speed results and identifying similar trips. Fine-resolution road network data constitute the primary input of such a network partitioning process. However, the spatial resolution of the analyzed UADP data determines the minimal resolution of the regional segmentation of the city. For instance, the spatial resolution of cell phone data generally corresponds to the underlying base station network. In this case, the partitioning of the urban network must result in larger regions than the Voronoi tessellation of the base station network.

The regional partitioning of the network impacts the data structure required for several inputs.

On the one hand, our method requires that the vehicle trips database (the key input of the method) include a coarse representation of the trip trajectory consistent with the previously defined scale, called a regional path. Therefore, the network partitioning affects this feature of vehicle trips. The other trip features are the observed arrival time and observed travel time. This minimal travel data structure corresponds to a generic intermediate format that is reasonably accessible by preprocessing the raw UADP data, regardless of their specific characteristics. The implementation details of this preprocessing step depend on these specific characteristics. They are not addressed here to preserve the generality of our framework. However, in Section 5.5, we shed light on the challenges linked to this step and provide options to overcome them.

On the other hand, the regional partitioning also constraints the average trip length estimates matrix, a critical input for the speed calculation phase. This matrix records local average trip lengths according to different macroscopic itineraries. Section 5.2.3 provides more details on its structure. This matrix is computed once, offline, and before the trip data analysis. This calculation can be based on the analysis of GPS data, if available, as done in this study. Those data must have sufficient coverage to calculate statistically reliable distances. However, alternatives exist, such as methods that exploit travel surveys or the automatic and systematic analysis of the road network topography [Batista et al., 2019].

Finally, the travel time estimation step requires an accurate evaluation of the average travel time bias caused by uneven user activity patterns. This evaluation, which relies on an analysis of the specific UADP data, is considered to be an input of the method. It will allow the observed travel times to be de-skewed and the average travel times to be estimated correctly.

Figure 5.1 illustrates the succession of the methodological steps and their articulation with the different inputs cited above. The following sections describe in more detail each of these steps.

### 5.2.3 Network partitioning and time resolution definitions

One of the essential steps in the methodology is the identification and the fusion of similar trips. However, sparse trips distributed over space and time are difficult to compare and relate to one another. In this section, we first propose to define a new spatial and temporal scale, thus laying the ground for the definition of comparison criteria between different trips. The definition of such a new scale relies on both spatial and temporal aggregation.

We first define a new spatial scale. The targeted urban road network is partitioned into regions. These regions must mainly be characterized by homogeneous city fabric, demography, road network topology, and, most importantly, traffic dynamics. Homogeneity of traffic dynamics is an essential requirement for a robust estimation of the regional mean speed, as shown by the literature on the Macroscopic Fundamental Diagram [Daganzo, 2007, Geroliminis and Daganzo, 2008]. Following the network partitioning guidelines provided by the related literature, one can divide a city into a set of regions usually ranging from 5 to 20. This new spatial scale will later determine the final spatial resolution of traffic speed estimates. Therefore, it must be adapted to the precision requirements of the case study and, where appropriate, to the resolution of the data, as mentioned above. This regional scale provides the background for defining a fundamental notion of our method, the *regional path*:

- The *regional path* is defined as the sequence of the successive regions traveled from the origin to the trip destination. Therefore, it is a coarse representation of the path followed at the road segment scale, consistent with the regional partitioning of the network.

This up-scaling process is illustrated in Figure 5.2. While Figure 5.2a displays an individual trip at the road segment network, Figure 5.2b represents its corresponding regional path $R_1R_4R_6$.

Further on, we will consider that trips follow the structure defined here:

- We call *trip* the ternary structure defined by a regional path, an observed travel time and an observed arrival time.

The trip length estimation that must be performed beforehand of the method is also constrained by the previously defined regional scale and paths. This input shall record the average regional trip length in each region along each possible regional path. Thus, it can take the shape of a distance matrix $\hat{\boldsymbol{L}}$ where rows are the different possible regional paths, and columns are the different regions resulting from the spatial tessellation. The cell value at $(i, j)$ corresponds to the average distance traveled in the $j^{th}$ region, when traveling along the $i^{th}$ regional path $P$. It is equal to zero if the path $P$ does not cross the $j^{th}$ region. This matrix is assumed to be constant over time, but time-dependent patterns can be considered if they can be characterized independently on another dataset [Batista et al., 2021a].

Besides the change of spatial scale, we define a new temporal resolution. The evaluation period is discretized into equal time intervals. This new temporal reference imposes the temporal granularity of the speed evaluation and must be chosen accordingly. In particular, the temporal unit must be small enough to reproduce the rapidly changing speed dynamics during peak hours. We choose 15 minutes in this study, as commonly used in the literature.

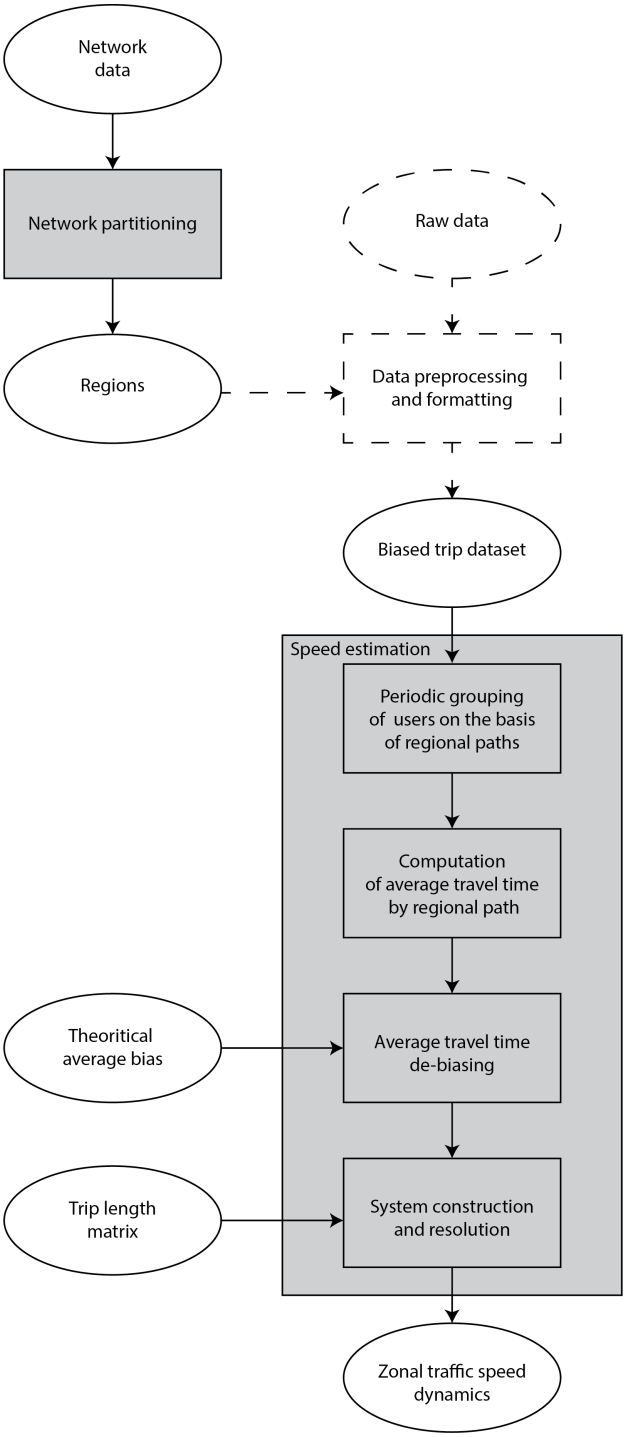
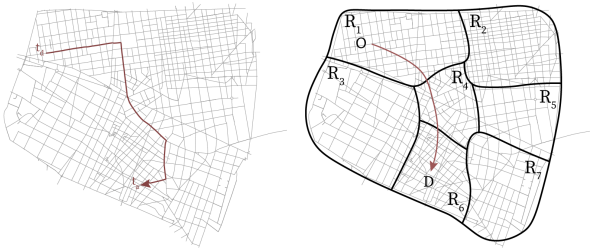



Figure 5.1: Methodological framework




(a) (b)

Figure 5.2: Representation of the different data quality for a same individual trip. (a) GPS track of an individual departing from their origin at time $t_d$ and arriving at time $t_a$. (b) The scaling up of the track at a regional scales accounts for those inaccuracies and reduce the route to a core path feature: the regional path $R_1R_4R_6$

These processes of partitioning the temporal dimension and the studied network, and the resulting notion of the regional path, provide temporal and spatial criteria for comparing different paths that are otherwise difficult to compare and the basis for identifying similar trips. The following relations are defined:

- $R1$: Two trips that share the same regional path are called **spatially similar**.
- $R2$: Two trips that share the same (exact) arrival period are called **simultaneous**.
- $R3$: Two trips that satisfy both $R1$ and $R2$, *i.e.*, that are **spatially similar** and **simultaneous**, are called **overlapping**.

These rules allow establishing a comparison between individual trips. This comparison is a crucial aspect of our methodology, which relies on identifying overlapping trips and fusing their observed travel time before de-skewing it.

However, as previously anticipated, it is essential to remind the difference between the exact arrival time and the observed one (extracted from UADP data). Such travel times might correspond to different periods if the user's communication rates are low. Therefore, identifying overlapping trips theoretically requires correcting the observed biased arrival time. Nevertheless, in the two following sections, we neglect this bias and consider that the exact arrival time is known when presenting the core methodological framework. We will relax this favorable assumption in Section 5.2.6.

### 5.2.4 Average travel time estimation

The robust estimation of travel times over the network is a critical milestone in our methodology. The travel time observations featured in the trip data can provide, to some extent, a snapshot of the traffic conditions that individuals encounter along their regional route at a given period. However, at an individual level, these observations are not reliable enough because they are sensitive, on the one hand, to the microscopic origin, destination, and routing of trips at the network level and, on the other hand, above all, to the frequency of individual observations.

As anticipated in the previous section, the observed travel time of a trip can be related to its exact travel time via the introduction of an additive temporal bias. Although other (non-additive) forms of bias could be considered, this model is the simplest to start with, and *a priori* the most natural. Let $P$ be a regional path, and let $i$ be an individual trip traveling along $P$. We thus have:

$$T^i_{P,obs} = T^i_P + \epsilon^i \tag{5.2}$$

where $T^i_{P,obs}$, $T^i_P$ and $\epsilon^i$ are, respectively, the observed travel time of $i$ along $P$, the exact travel time of $i$ along $P$, and the travel time bias of trip $i$.

Although estimating this individual bias would allow de-skewing the observed travel time, this bias is, by nature, difficult to assess. However, the estimation of its average seems less challenging and can allow to de-skew on average the observed travel times. This average bias is assumed known and to be an input of our framework. This hypothesis is discussed in Section 5.5. To this end, we propose merging overlapping trips and averaging their observed travel times to build a unique aggregated biased travel time information by path and period.

Let $t$ represent a generic period, and let $I^t_P$ be the set of overlapping trips along P that reach destination at time $t$, with $n_{t,P} = |I^t_P|$. Averaging Equation 5.2 over $I^t_P$ gives:

$$\bar{T}^t_P = \bar{T}^t_{P,obs} - \bar{\epsilon}^t_P \tag{5.3}$$

where $\bar{T}^t_P$, $\bar{T}^t_{P,obs}$ and $\bar{\epsilon}^t_P$ are, respectively, the average actual travel time, the average observed travel time, and the average bias of trips from cluster $I^t_P$.

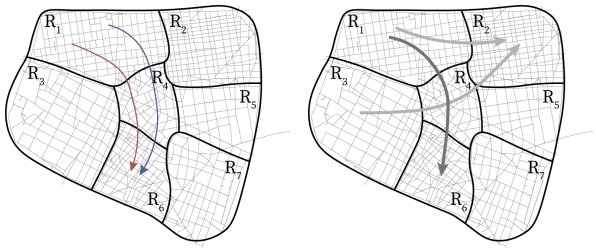
Assuming that the bias is independent of the trip path and time (hypothesis $H_2$, discussed below), we can consider that the distribution of individual biases $\epsilon^i$ can be modeled via a unique random variable $X$. The construction of such a model, and the estimation of its first moment $\mu_X \equiv E(X)$, can offer an approximation of $\bar{\epsilon}^t_P$ allowing the de-skewing of $\bar{T}^t_{P,obs}$, provided that the sample of individuals associated with this period and path is large enough:

$$\bar{T}^t_P \approx \bar{T}^t_{P,obs} - \mu_X \tag{5.4}$$

One of the great advantages of merging overlapping trips data together is that it makes the estimation of the average travel time more robust, as long as $\mu_X$ can be independently characterized. This trip aggregation greatly reduces the complexity of estimating travel times and traffic speed from biased temporal data, since neither the estimation of the individual biases, nor the characterization the bias distribution are needed. Only the estimation of its average value is required. However, the sampling size is a key condition of the process: the larger the sample is, the better the theoretical average bias $\mu_X$ is representative of the sample's average bias.

The temporal independence of the bias can be discussed in light of the work of Chen et al. [2018], who showed, using CDR data, that the inter-event time distribution is sensitive to the hour of the day and. In particular, longer inter-event times are observed during nighttime and early morning. However, these results account for all individuals, including the ones that are static and sleeping, while we are exclusively interested in moving ones. Thus, our hypothesis comes down to considering that users' communication activities are more related to their general activity level (mobile or static) than to the hour of the day, which seems reasonable. The spatial independence of the data is similarly debatable since mobile phone or social network use is correlated with socio-demographics. Thus, it would be interesting to validate or refute our hypothesis with a study of the evolution of the inter-event time of traveling users through time and space, but this goes beyond the scope of this chapter. The assumption made here allows considering a first simple de-skewing approach. Future researches on the travel time bias associated with UADP data could further complete this approach by differentiating the average bias according to time or space.

The systematic estimation, for each $P$ and $t$, of the observed travel times, and their de-skewing using an average bias estimate results in a robust, spatially exhaustive, and dynamic evaluation of the travel times across the network at each period. In that sense, Figure 5.3 illustrates how two overlapping trips are jointly analyzed to build a unique representative object of the traffic conditions along $R_1R_4R_6$. Figure 5.3b also shows how




(a) (b)

Figure 5.3: Visualization of the data merging into clusters of similar trips. (a) Two individuals traveling simultaneously along a same regional path despite following different (unknown) routes. (b) Merging those individuals into a unique average object (in dark grey) allows characterizing the average travel time needed to travel the regional path $R_1R_4R_6$. This is repeated for every regional path and helps in characterizing the travel time over the whole network.

this can be repeated for every regional path of the network. The speed estimation process relies on this systematic mean regional path travel time estimation.

### 5.2.5 Speed estimation

This section develops the mathematical foundations of the speed estimation method.

Starting at the individual level, we consider an individual trip $i$ of $I_P^t$. Its exact traveled time $T_P^i$ along $P$ can be expressed as the sum of the traveled times $T_{P,r}^i$ over each region $r$ of $P$ (see Equation 5.5). The regional travel time terms can be in turns expressed as the fraction of the distance traveled by $i$ in $r$ (i.e., $L_{P,r}^i$) over the mean spatial speed of $i$ in region $r$ (i.e., $V_r^i$), as described in Equation 5.6.

$$\forall i \in I_P^t, \qquad T_P^i = \sum_{r \in P} T_{P,r}^i \tag{5.5}$$

$$T_P^i = \sum_{r \in P} \frac{L_{P,r}^i}{V_r^i} \tag{5.6}$$

Due to the data temporal sparsity of individual tracks, $T_{P,r}^i$, $L_{P,r}^i$, and $V_r^i$ are considered unknown.

Although vehicles may experience different local and instantaneous speeds over an area, their average speeds depend mostly on overall traffic conditions, and mainly on the accumulation (*i.e.*, number of vehicles in the region). These speeds show little scatter among individuals, and can be approximated by the mean spatial speed of all individuals traveling in the region. This observation has sustained the development of the MFD theory [Daganzo, 2007, Geroliminis and Daganzo, 2008]. The partitioning of the network into subregions of consistent traffic dynamics is especially meant to enforce this assumption. On this basis, we assume that each regional speed is homogeneous and constant over the duration of each period $t$, so that:

$$V_r^i = V_r^t, \quad \forall i \tag{5.7}$$

where $V_r^t$ is the regional spatial mean speed at period $t$.

In Equation 5.6, after summing on the $I_{t,P}$ trips, this gives:

$$\sum_{i=1}^{n_{t,P}} T_P^i = \sum_{i=1}^{n_{t,P}} \sum_{r \in P} \frac{L_{P,r}^i}{V_r^t} = \sum_{r \in P} \sum_{i=1}^{n_{t,P}} \frac{L_{P,r}^i}{V_r^t} \tag{5.8}$$

Equation 5.8 can easily be rewritten as follows:

$$\sum_{i=1}^{n_{t,P}} T_P^i = \sum_{r \in P} \frac{1}{V_r^t} \sum_{i=1}^{n_{t,P}} L_{P,r}^i \tag{5.9}$$

$$n_{t,P} \bar{T}_P^t = \sum_{r \in P} n_{t,P} \frac{\bar{L}_{P,r}^t}{V_r^t} \tag{5.10}$$

$$\bar{T}_P^t = \sum_{r \in P} \frac{\bar{L}_{P,r}^t}{V_r^t} \tag{5.11}$$

Again, a significant advantage of this averaging process over the sample $I_P^t$ is that the characterization of individual regional trip lengths $L_{P,r}^i$ for any individual $i$ becomes unnecessary. Instead, the sample mean value $\bar{L}_{P,r}^t$ turns out to be sufficient. On condition that the sampling size is large enough, this can be replaced by its static estimate $\hat{L}_{P,r}$, drawn from the exogenous trip length matrix $\hat{\boldsymbol{L}}$ described above:

$$\bar{T}_P^t \approx \sum_{r \in P} \frac{\hat{L}_{P,r}}{V_r^t} \tag{5.12}$$

At this stage, the computed distance matrix is used to express, through Equation 5.12, a relationship between the average travel time along path $P$ at period $t$, and the underlying, unknown traffic speeds of the regions along the $P$.

Although the average trip duration $\bar{T}_P^t$ is unknown, in Section 5.2.4 we discussed how a knowledge of the average time bias $\mu_X$ could allow to estimate it. Based on Equation 5.4, we thus get:

$$\bar{T}_{P,obs}^t - \mu_X \approx \sum_{r \in P} \frac{\hat{L}_{P,r}}{V_r^t} \tag{5.13}$$

At each period $t$ and for each path $P$, the average travel time along $P$, $\bar{T}_{P,obs}^t$, can be derived from the UADP analysis. Conversely, the constant distance parameters $\hat{L}_{P,r}$ are drawn from the aforementioned estimated trip length matrix $\hat{\boldsymbol{L}}$. $\mu_X$ is assumed known as well. $V_r^t$ are the only unknowns of the system. When applying in Equation 5.13 the change of variable $x_r^t = 1/V_r^t$, we finally get the unbiased system:

$$\forall t, \quad S^t = \{\bar{T}_{P,obs}^t - \mu_X = \sum_{r \in P} \hat{L}_{P,r} x_r^t, \quad \forall P\}. \tag{5.14}$$

In Equation 5.14, we name $S^t$ the linear system composed of $|R|$ unknowns $(x_r^t, r \in R)$ and as many equations as the number of regional paths observed during the reference period $t$. The UADP data analysis and the parameters extracted from the trip length matrix allow to fully characterize the system, which can be rewritten in matrix notation as:

$$\forall t, \quad S^t = \{\boldsymbol{T_{obs}^t} - \boldsymbol{\mu_X} = \hat{\boldsymbol{L}}^{|\boldsymbol{t}}\boldsymbol{x^t}\} \tag{5.15}$$

where $\boldsymbol{T_{obs}^t}$ is the average observed travel time vector and $\hat{\boldsymbol{L}}^{|\boldsymbol{t}}$ is the sub-matrix of $\hat{\boldsymbol{L}}$ restricted to the regional paths observed at period $t$.

Given that the number of regional paths will generally exceed the number of regions of the adopted partitioning, $S^t$ is very likely over-determined. Consequently, the system will probably have no exact solution, but an approximated one can be calculated using regression analysis. To this purpose, we apply a non-negative least squares regression method to the system. For a given over-determined linear system $\boldsymbol{Ax} = \boldsymbol{y}$, in which $\boldsymbol{A}$ is a matrix, $\boldsymbol{x}$ the unknown vector and $\boldsymbol{y}$ the response one, the ordinary least square problem consists of finding the optimal $x$, which minimizes the sum of the squared residuals. This can be formulated as solving $\boldsymbol{x_0} = argmin_{\boldsymbol{x}}||\boldsymbol{Ax} - \boldsymbol{y}||_2$, with $||.||_2$ the euclidean norm. Additional constraints on the elements of $x$ can be added. This is the case in the non-negative least square method, implying that the coefficient of $\boldsymbol{x}$ be non-negative. In our case, such constraint allows for taking into account the non-negative nature of zonal traffic speed. We apply the non-negative least square method to $S^t$, by solving at each time step the following:

$$\boldsymbol{x_0^t} = argmin_{\boldsymbol{x}}||\hat{\boldsymbol{L}}^{|\boldsymbol{t}}\boldsymbol{x^t} - \boldsymbol{T_{obs}^t} + \boldsymbol{\mu_X}||_2, \quad \boldsymbol{x} \geq 0 \tag{5.16}$$

The non-negative least square method *nnls*, implemented in Python's package *Scipy*, was used here. Taking the reciprocal values of the solution vector $\boldsymbol{x_0^t}$ gives the optimal speed vector $\boldsymbol{v_0^t}$. This resolution process can be iterated throughout the whole studied time span to estimate the complete temporal speed trends. It should be noted the intra-region trips were filtered out of the system and discarded from the analysis, as they contribute to a diagonal subsystem whose optimization seems to take precedence over the other system equations in the regression analysis.

### 5.2.6 Arrival time correction and data selection

The previous sections have considered the arrival period $t$ to be exact. However, when extracting trips from UADP data, not only the travel time is biased, but so are the arrival time and period. Let $t_0^i$ be the actual precise arrival time, and $t_{0,obs}^i$ the observed precise arrival time, by opposition to $t^i$ and $t_{obs}^i$ that refer to the actual and observed arrival periods. We have:

$$t_{0,obs}^i = t_0^i + \epsilon_a^i \tag{5.17}$$

When reducing the temporal resolution to the period level, this implies that the observed arrival period does not necessarily corresponds to the actual arrival period. This results in the following inequality:

$$t_{obs}^i \geq t^i \tag{5.18}$$

Consequently, identifying simultaneous trips based on the observed arrival period might correspond to considering together users that refer in reality to other periods, with potentially different traffic speeds. Therefore, the correct gathering of simultaneous trips ideally requires recovering for each individual their exact arrival times from their observed arrival

times. This recovering cannot be done on average, as for the travel time de-skewing. Deducting the expected value of the arrival bias (*i.e.*, half of the expected value of the travel time bias $\mu_X$) from the observed arrival times of each trip, as in Equation 5.19, only shifts all trips by the same amount of time, but not re-assign each trip to its correct arrival period.

$$t_0^{i\prime} = t_{0,obs}^{i} - \mu_X/2 \tag{5.19}$$

While this shift may help correct an average time offset and slightly modify individuals' grouping, it can in no way correct the massive mixing of trips together. Such a correction requires the precise estimation of individual biases separately, which seems very hard to achieve considering the nature of the data. Therefore, we abandon the idea of applying such an individual bias correction and stick to correcting the average arrival time offset by considering the new arrival period $t_0^{i\prime}$ as defined by Equation 5.19. Nevertheless, to fully meet with the challenge raised by this arrival bias, we also propose to enhance at each period the robustness of the linear system by implementing filtering solutions at different levels.

Firstly, one can consider filtering individuals according to a criterion based on their communication rates, in order to limit to some extent the mixing of individual trips corresponding to different periods. In practice, the individual overall average inter-event time can be used as an indicator of these communication rates. However, this filter must be considered with caution, as it might impact the sampling size. In our study, trips are not associated with individual inter-event times but with individual biases. We will explore the impact of filtering trips based on a criterion addressing these biases.

Second, we suggest implementing a filter on the minimal number of trips to consider that an equation defining a regional path at a given period is valid. Setting this minimal threshold aims at ensuring the robustness of the travel time estimates despite potential shuffled trips. One could also consider setting a maximum threshold on the travel time standard deviation, which could be particularly suitable for large trip samples.

The criterion above focuses on the reliability of each equation independently of others. A third element we consider is the consistency of the equations with each other. As this coherence is quite challenging to evaluate, we propose a sensitive filtering approach to stabilize the results obtained from a set of indiscriminate equations. The approach we propose is based on bootstrapping, a statistical inference method based on random sampling. For a given period, for which the data processing resulted in a system $S$ made of $n$ distinct equations, a set of subsystems $S_i$ is generated and solved to explore the sensitivity of the results to the structure of the system. Specifically, the generic subsystem $S_i$ is built by sampling with replacement the same number $n$ of equations from $S$. Consequently, $S_i$ has the same number of $n$ equations but possible redundancy for some of them. To take this redundancy into account, we resolve the system with a weighted least square optimization method. The weight of each equation is given by its number of occurrences in the subsystem. Thus, the more an equation is sampled from the original system, the higher its weight in the resolution. This process is iterated over many subsystems (we set the minimal number of iterations to 100) to explore the results' sensitivity to different sampled equations and weighting parameters. Consequently, many derived speed solutions are generated at each period, resulting in a speed distribution for each region. We apply statistical filters to these distributions to filter out the most aberrant values before averaging the remaining speeds. This process enforces the results' consistency and stabilizes them without explicitly labeling equations as reliable or not and arbitrarily filtering them out.

Although our method does not exclude does not exclude working with a favorable subpopulation displaying the lowest communication inter-event times, the individual filtering

limits the reach of the method by reducing the range of users considered. Filtering individuals according to their inter-event travel time corresponds to considering a sub-population with a reduced average bias. Therefore, among these three filtering methods, the last two are considered preferable in the evaluation of our methodology.

### 5.2.7 Speed trends smoothing

The speed estimation process described above is applied independently at each time step. The results of this recursive application of the method to consecutive periods may present sawtooth instabilities between consecutive periods due to variations of the regional paths observed, their number or the amount of travelers they represent. To smooth speed trends over time and ensure consistency of results between consecutive periods, we implement a dynamic filtering based on a rolling window method. The window size is set to a chosen number of periods $n$. At each period $t$, the smoothed traffic speed is calculated as :

$$\bar{V}_r^t = \frac{1}{n} \sum_{i=0}^{n-1} V_r^{t-\frac{n-1}{2}+i} \tag{5.20}$$

Because the speed trends can vary faster at peak hours than during the remaining periods of the day, we increase the sensibility of the filter at this time. Then, $n$ is set to 3 (periods) during assumed peak hours, while it is set to 5 the rest of the day.

### 5.2.8 Discussion

In this section, we discuss a few insights we can retrieve from the structure of the system.

First, the structure of the system $S^t$ directly explains the impact of the chosen tesselation. The more fine-grained the spatial resolution, the larger the system size. Not only does the number of unknown variables (regional speeds) increase, but so does the number of possible regional paths, and hence of equations. Consequently, an increase in the number of regions also leads to a relative decrease in regional paths' attendance level, as they are more numerous and therefore less crossed. This attendance decrease might be problematic as the methodology relies on the hypothesis of sufficient sample representativeness. Thus, determining the appropriate spatial partitioning raises the question of finding the suitable trade-off between a fine-grained traffic speed estimation and a system composed of reliable equations.

Additionally, the shape of the system provides an insight into the importance of the average travel time de-skewing. The speed vector resulting from the approximated resolution of the system $S^t$ of Equation 5.15 is reliable under the condition that the system is properly conditioned, *i.e.*, that the average travel time vector $\boldsymbol{T_{obs}^t} - \boldsymbol{\mu_X}$ is correctly estimated (the trip length matrix distance factors $\hat{\boldsymbol{L}}$ being considered as reliable). Without accounting for the average bias generated by the users' uneven activity rhythms, characterizing the regional network travel times based on the observed travel times would result in an overstated left side of the system compared to the latent actual average travel time $\bar{T}_P^t$. This system would be unrepresentative of the actual traffic speeds and likely to underestimate them.

## 5.3 Experimental approach

To evaluate the performance of the proposed methodology, we apply it to a UADP dataset derived from high-frequency GPS data through data simplification and downsampling. This evaluation approach, based on high-frequency raw data instead of low-frequency data,

presents several advantages. First, it provides control over the average data bias, which is an essential part of the methodology that has not been enough characterized by the literature. Second, it allows exploring the impact of the data simplification, the temporal downsampling, and the de-skewing process on the speed estimation quality. This exploration is a necessary step to identify the strengths and weaknesses of the method and adjust it accordingly. It helps to understand how to increase the robustness of the method before applying it to real inaccurate and biased UADP data, for which the corresponding accurate ground truth GPS tracks will most likely be lacking. Last but not least, this experimental approach provides an easily accessible and reliable baseline estimation of the traffic speed dynamics, against which to compare our results. The original GPS dataset also provides valuable data for estimating the trip lengths, which are assumed in our framework to be derived from exogenous sources and produced offline before UADP-like data processing.

### 5.3.1 Bias model

In the previous section, we discussed how the temporal biases of the data could substantially impact the speed estimation results and why it was necessary to take into account this bias in order for the speed estimation system to be adequately conditioned. This led to Equation 5.15. To the best of our knowledge, no model is characterizing this bias. Thus, we propose a simplistic and generic bias distribution modeling to downsample GPS trips and simulate the temporal characteristics of data with low-sampling rates. The modeled bias is positive and independent both of time and space, in agreement with the considerations and assumptions developed in Section 5.2. The model relies on the characterization of the sample inter-event time distribution, which is a standard indicator to measure the sampling rates of user-activity-dependent positioning data. The generated bias distribution will be applied to the individual trips in order to simulate temporal biases and explore the performance of the method on these downsampled trips.

The travel time bias is modeled by a random variable X, which we aim to characterize. To start, we express the individual travel time bias $\epsilon^i$ as the sum of a departure and arrival offsets $\epsilon^i_d$ and $\epsilon^i_a$ (see Equation 5.1):

$$\epsilon^i = \epsilon^i_d + \epsilon^i_a \tag{5.21}$$

$\epsilon^i_d$ being the time difference between the actual and observed departure times from the origin stay, and $\epsilon^i_a$ the time difference between the observed and actual arrival times at the destination stay. They are considered to be positive. Hence, if the departure and arrival temporal offsets $\epsilon^i_d$ and $\epsilon^i_a$ are themselves modeled by the same random variable Y, Equation 5.21 gives $X = 2Y$. Now, we consider that an individual's departure time from a stay position can occur with a uniform probability between the pre-departure communication event and the post-departure communication event. The delay between these two events follows the distribution of the user's inter-event times, which we assimilate with the population's inter-event times distribution for the sake of simplicity.

Mathematically, this means that departure bias follows a uniform distribution law bounded by the population's inter-event time distribution. Symmetrically, the same reasoning applies to the arrival bias.

Here, we model the population's inter-event time by a simple exponential law $Z$ of parameter $\lambda$. While the literature often reports inter-event time distribution closer to truncated power law distribution, we select an exponential distribution here out of simplicity. It requires a single parameter $\lambda$ directly linked to the distribution average.

The considerations above lead to:

$$Z \sim Exp(\lambda) \tag{5.22}$$
$$Y|Z \sim U(0, z) \tag{5.23}$$

Hence, the probability density function of Z, and the conditional probability density function of $Y$ given the occurrence of the value $z$ of $Z$ can be written as:

$$f_Z(z) = \lambda e^{-\lambda z} \tag{5.24}$$
$$\text{and } f_{Y|Z}(y \mid z) = \begin{cases} \frac{1}{z} & 0 \leq y \leq z, \\ 0 & \text{otherwise,} \end{cases} \tag{5.25}$$

From this, we show (see the detailed calculation in Appendix C.2) that the probability density function of $Y$ is:

$$f_Y(y) = \lambda \int_0^{+\infty} \frac{e^{-\lambda(y+z)}}{y+z} dz \tag{5.26}$$

and that the first two moments of Y are:

$$E(Y) = \frac{1}{2\lambda} \qquad \text{and} \qquad V(Y) = \frac{5}{12}\frac{1}{\lambda^2} \tag{5.27}$$

Those results characterize the random variable $Y$ which models the departure and the arrival offsets. This gives for $X = 2Y$:

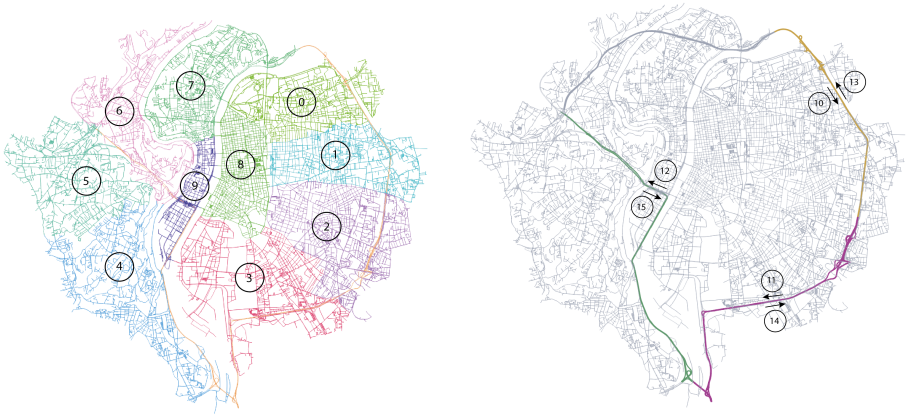
$$\mu_X \equiv E(X) = 2E(Y) = \frac{1}{\lambda} \qquad \text{and} \qquad V(X) = 4V(Y) = \frac{5}{3}\frac{1}{\lambda^2} \tag{5.28}$$

We already stressed the importance of the size of $I_P^t$ to ensure that $\mu_X$ is representative of the cluster average bias. This is all the more important as with our bias model, as the variance of X increases with mean inter-event time $E(Z) = \frac{1}{\lambda}$:

$$V(X) = \frac{5}{3}E(Z)^2 \tag{5.29}$$

Equation 5.29 shows that the larger the mean inter-event time is, the more scattered the trip bias distribution will be, and the more data per period and per regional path will be needed to ensure a reliable de-biasing process.

With this model, fairly simple and realistic, we propose a way to simulate the travel time biases related to the users' variable mobile phone activity rates. The construction of the model makes it possible to approximate the average bias of the measured travel times from the analysis of the population inter-event time distribution and deduce it from them. While an exponential inter-event time distribution was chosen here, the method is transposable to any other observed distribution. Although this model was developed for the data simulation purposes and not validated against real UADP data, we look forward to evaluating its relevance in a real context. We also believe that the relationship between bias and inter-event time will remain a key part of a more complex bias modeling process.




(a) (b)

Figure 5.4: Maps of the regions partitioning the city of Lyon, France. (a) Map of the urban regions; (b) Map of the ring road regions. The ring road is divided into three zones, which are themselves separated into two according to the direction of traffic.

### 5.3.2 Spatial partitioning

The city of Lyon, France, is chosen as our case study. The study area includes Lyon and the neighboring municipality of Villeurbanne, located inside Lyon's ring road. We have parted this territory into sixteen distinct regions. Ten of them divide the urban areas, while the ring road is extracted and parted into six regions, three per each direction. Those regions are displayed in Figure 5.4. Urban regions were manually defined based on the natural geographical barriers (two rivers) and the major road networks. The major adopted criteria consisted of separating the main arterial roads in different regions. The traffic variables were verified to be relatively uniform in each region [Mariotte et al., 2020]. We split the ring road into three main blocks based on our knowledge of daily congestion patterns: the north-east, south-east, and south-west blocks. The remaining north-west section of the ring road is mainly a tunnel. As the GPS data are lacking in this section, it was ignored in the analysis. We checked whether the two opposite travel directions could be jointly considered a homogeneous traffic area by analyzing the ring road speed profiles. As the speed profiles appeared to be significantly different, we decided to split the ring road further, regions per direction. It is important to mention here that despite the efforts to ensure the homogeneity of the traffic conditions inside each zone, some aspects of the network structure can be a limitation. In particular, many motorways serve Lyon and relate it to the neighboring cities. Those motorways cross the urban regions and cause within regions traffic heterogeneity (region 0, 2, 3, 4, and 6). One solution to limit this heterogeneity would be to isolate those motorways sections into new specific regions. However, this would unnecessarily increase the number of traffic speed variables. Instead, we propose a light and easy-to-implement adaptation of the overall methodology to take this aspect into account. Although this filter is specific to our case study and the chosen partitioning, it can be applied again in other contexts, as cities are often served by expressways passing through peripheral residential areas.

Based on our knowledge of the traffic in Lyon, we assume that the trips traveling along those motorways are very likely to travel along the ring road as well, as a transition to another motorway or their final destination in an urban region of the city. Consequently, the ring road is assumed to be more strongly connected to these motorways than to the rest of the urban regions. Hence, we propose to decouple our estimation equation system as follows.

On one side, a first subsystem $S^t_{RR}$ is built from the regional paths that travel along

the ring road at one point. The system is solved and returns a first speed vector $V_0^t$. The corresponding equations are assumed to carry reliable information about the traffic speed on the ring road. However, the information they carry about the dynamics in the other urban regions (traveled before or after the ring road) is assumed to characterize better the traffic condition in their motorways than in their urban grid. Consequently, while the solution $V_0^t$ is considered reliable for characterizing the ring road speed, it is considered as unreliable to characterize the urban regions' speeds.

On the other side, we build a second subsystem $S_{URB}^t$ with regional paths that do not travel along the ring road. The resolution of this new subsystem results in a second speed vector $V_1^t$. This solution only characterizes the urban regions and is assumed to be reliable on them.

Both solutions are merged to build a unique speed vector $V^t$ built from the concatenation of $V_{1|URB}^t$ the speed vector $V_1^t$ restricted to the urban components and $V_{0|RR}^t$ the speed vector $V_0^t$ restricted to the ring road components.

### 5.3.3 Data description

The GPS dataset exploited in this study consists of cleaned and map-matched GPS traces over the Greater Lyon area, i.e., an area larger than the perimeter selected for our study. A European navigation system provider collected the data between October 2017 and September 2018. The traces are collected from multiple navigation system technologies equipping a multitude of observed floating vehicles (29,000 vehicles per day on average on the Greater Lyon). Moreover, as each trace corresponds to a vehicle, there is no need to filter out pedestrian or cyclist travelers as usually required when working with mobile phone or social networks data. This aspect slightly facilitates the problem of estimating traffic speeds, since the question of detecting the mode of transport does not arise here.

The trips used in this study were extracted from five typical weekdays, i.e., from Monday, February 12, to Friday, February 16, 2018. As few trips are observed at night-time in our dataset, the time span selected for our evaluation is restrained to day-time hours, *i.e.*, in-between 5 AM and 8 PM. The data from the full month of February 2018 was used for the offline calculation of trip lengths.

### 5.3.4 Trip data preparation

The first phase of data processing involves filtering and further cleaning the data. As the area covered by the GPS data is larger than the studied perimeter, we applied a first filter to remove from the data the segments of GPS tracks outside the relevant perimeter. Moreover, the GPS tracks are additionally parsed into different trips when stays are detected. Additional steps included filtering out redundant individuals, static vehicles, and GPS tracks that are fragmented or do not have a spatial consistency, to obtain a clean and reliable data set. At the end of this preprocessing step, the number of trips per considered day is as described in Table 5.1.

Although these numbers are significant, we insist on the importance of a minimal sample size at the period and regional path level. At each time step, and for each path, the number of trips must be large enough so that the expected value of the bias is representative of the sampled biases. As GPS data are limited in sample size, we artificially extend the size of the dataset by duplicating each trip 100 times. This trick allows obtaining an extended sampled population, that is then downsampled and biased for each individuals.

This GPS trip dataset is then strictly reduced to the trip features needed by the methodology. The actual travel and actual arrival times of each trip are directly extracted from the GPS data observation. Additionally, every GPS track is down-scaled to the spatial resolution previously defined, to obtain the regional path information. Those three trip

| | Day 1 | Day 2 | Day 3 | Day 4 | Day 5 |
|---|---|---|---|---|---|
| Number of trips | 19597 | 20750 | 20951 | 21963 | 22302 |

Table 5.1: Number of trips considered per day

features (*regional path*, *actual arrival time*, and *actual travel time*) are stored, along with the *trip id*, in a new dataset that will be called $DS_0$ in the following. At this stage, a first downsampling level has been introduced in the spatial dimension to replace the precise track information with the regional path feature. Although travel times do not yet include any temporal bias at this stage, the trip representation is then already considerably simplified. This dataset will be the subject of our first experiments.

The last processing step consists of applying to $DS_0$ a temporal downsampling process that aims to simulate the temporal imprecisions of UADP data compared to GPS data. The idea is to simulate the travel time increase caused by the temporal biases that the uneven inter-event communication times introduce in the departure and arrival time detection, using the bias model described in Section 5.3.1. The average inter-event time (IET) is a crucial parameter of this model. The value of this parameter may depend on the population observed or on the type of data chosen: for example, using handovers and signalization datasets will display weaker inter-event times than CDR or LBNS data. To take this inter-event time variability into account and evaluate its impact on our results, we generated, for each day of data, five different downsampled dataset, one per inter-event time value. The selected average inter-event time values are 4, 8, 12, 16, and 20 minutes, to cover a large range of average communication rates. The corresponding bias distributions are displayed together in Figure 5.5. We observe that the larger the average bias is, the more spread is the distribution, with a greater probability for high temporal biases, which was expected with Equation 5.29. This plot allows understanding that even if the speed estimation method is statistically unbiased, the increasing dispersion of the individual biases makes it necessary to have larger samples when working with important average inter-event times compared to small ones. This especially justifies the data expansion led above. Downsampling the expanded trips sample then allows obtaining an extended bias distribution, for which the average bias will be more representative. For each trip, we sample the departure and arrival biases according to the probability density function obtained in Equation 5.26. We generate a second dataset, referred to as $DS_1$, which includes the same trips as in $DS_0$ but whose actual travel time information is biased by the sum of the sampled departure and arrival biases to obtain the observed travel time. This dataset records partially biased trips. It will be the subject of our second analyses to assess our ability to correct for travel time bias. In a final downsampling step, we generate a dataset $DS_2$ in which the actual arrival time is additionally biased with the sampled arrival bias. $DS_2$ records synthetic UADP-extracted trips: individual trips characterized by fully biased temporal features and low-quality path information.

One last step before applying the methodology to any of those two datasets consists of grouping the data by regional paths and 15-minutes periods, and averaging the travel time on the resulting groups.

### 5.3.5 Speed baseline

We divide the experiment duration into equal periods of 15 minutes. At each time step $t$, the method, applied to one of the datasets describe before, returns a vector $\boldsymbol{V}^t$ whose dimension is equal to the number of regions, in this instance 16. A speed reference is needed to validate our method and estimate the impact of the data downscaling and downsampling processes on the reliability of the results. The spatial mean speed $V_r^t$ in region $r$ over a

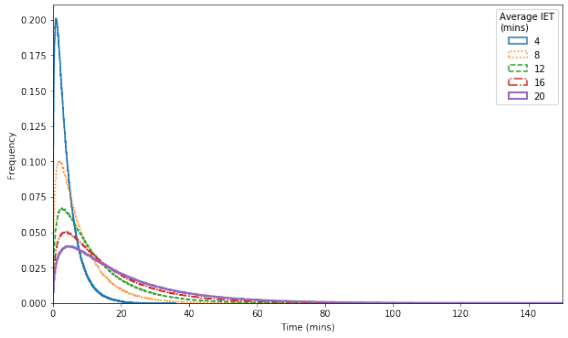



Figure 5.5: Bias distributions depending on the selected average inter-event time

period $t$ is defined as the ratio of the total traveled distance in region $TTD_{r,t}$ and the total travel time $TTT_{r,t}$ in region $r$ during $t$:

$$V_{r,t} = \frac{TTD_{r,t}}{TTT_{r,t}} = \frac{\sum_i d^i_{r,t}}{\sum_i t^i_{r,t}} \tag{5.30}$$

$TTD_{r,t}$ corresponds to the sum of the individual travel distances in region $r$ during $t$, i.e., $d^i_{r,t}$. $TTT_{r,t}$ corresponds to the sum of the individual travel times in region $r$ during $t$, i.e., $t^i_{r,t}$. As the GPS data are map matched (see Section 5.3.3), they include not only temporal and positional data, but also the inferred sequence of road segments traveled, the inferred entrance time on each road link, and the distance traveled on each of them. These characteristics give access to precise individual travel times and distances, and therefore allow obtaining a reliable speed baseline to compare speed estimation results from temporally-biased trip data.

To conclude, using simulated data obtained through downscaling and the downsampling of high-frequency GPS data presents the strategic advantage of offering substantial control over the experimental environment while providing easy access to the necessary distance parameters and the ground truth speed data. The next section exposes the results of the approach on datasets $DS_0$, $DS_1$ and $DS_2$.

### 5.3.6 Trip length estimation

Finally, the GPS data is used to estimate the trip length estimation matrix. The entire month of February is used to generate this matrix and ensure robust estimation of trip lengths. As the mobility behaviors are characterized by high redundancy, the trip lengths are mostly unvarying from one month to the other. This means that the trip length matrix can be calibrated using data from a period of time that does not necessarily overlap the time span of the study. This assumption was confirmed by comparing trip length matrices computed in February with a similar matrix computed with March's data. Appendix C.3 exposes the result of this comparison. We will explore the results of the method when estimating travel distance with automatic network analysis in later work. In the meantime, one can refer to the comparison of such trip lengths estimation with GPS data in the work of Batista et al. [2021b].

## 5.4 Results

The speed estimation method that we propose presents the advantages of relying on few mobility features and hence of being easily applicable to UADP data. However, it is

| | MAE (km/h) | RMSAE (km/h) | MAPE (%) | RMSAPE (%) |
|---|---|---|---|---|
| Day 1 | 4.727834 | 6.548142 | 13.111325 | 16.875215 |
| Day 2 | 4.820239 | 6.725782 | 13.169344 | 16.739542 |
| Day 3 | 4.906018 | 7.048133 | 13.486273 | 17.370823 |
| Day 4 | 4.781254 | 6.909102 | 14.909294 | 23.519969 |
| Day 5 | 4.876119 | 6.538737 | 14.013252 | 17.780414 |

Table 5.2: Daily speed errors when applying method to $DS_0$

essential to evaluate the extent to which the low data quality impacts the accuracy of the speed estimation. To this purpose, we proceed in three steps.

First, we intend to evaluate the impact of working with mobility data of coarser space and time resolution on the results by assessing the errors when working on the $DS_0$ dataset. The significant degrading of the GPS data might impact the results. Evaluating this impact is essential to understand the overall potential of the method on temporally-biased trip data.

In a second step, the method is applied to the partially biased dataset $DS_1$. First, we estimate the speed dynamics without de-biasing the temporal system, hence with erroneous travel time information, to evaluate to what extent it is necessary to estimate and remove the travel time bias. Second, we solve the de-biased system, and measure the effectiveness of the de-biasing process to obtain satisfactory speed results.

In a third step, we will apply the method to the fully biased dataset $DS_2$. We compare the results obtained when correcting only the travel time bias with the results obtained when correcting both the arrival and travel time biases. While the data expansion of our trip sample has no impact on the evaluation when using dataset $DS_0$, because it does not change the average travel times, we will see that this step is of importance when dealing with both biased datasets.

### 5.4.1 Method application to trip data with exact travel time

We start by applying the proposed methodology to dataset $DS_0$, to evaluate in a first step the impact of the spatial aggregation and of the speed estimation method.

By using our methodology, we obtain a speed profile in kilometers per hour, per region, and per 15-minutes slots for each day of the evaluation. These speed profiles are compared to the corresponding speed baseline to compute errors. We begin by measuring daily error indicators that characterize the global results of the methodology for the overall regions and periods of the day. We evaluate the mean absolute error (MAE), the root mean absolute error (RMSAE), the mean absolute percentage error (MAPE), and the root mean square absolute percentage error (RMSAPE). Those daily indicators are displayed in Table 5.2.

It is interesting to observe that the daily errors are substantially similar from one day to another.

To better assess the performance of the methodology, we now focus on one specific day from our day-set, *e.g.*, Day 1 (Monday, February 12, 2018). Comparable results were obtained for the other days. Figure 5.6 illustrates the speed estimations dynamics obtained for this day. Each of the subplots corresponds to a region of our partitioning of Lyon. Time throughout the day is represented on the x-axis in hours while the y-axis is for average traffic speed, in kilometer per hour. The ground truth traffic speed, calculated based on the raw GPS dataset, is represented in blue. The orange line corresponds to the raw speed estimation results after bootstrapping. The green line is the result of the moving average filtering that smooths the speed trends. The first ten plots (from Region 0 to Region 9)

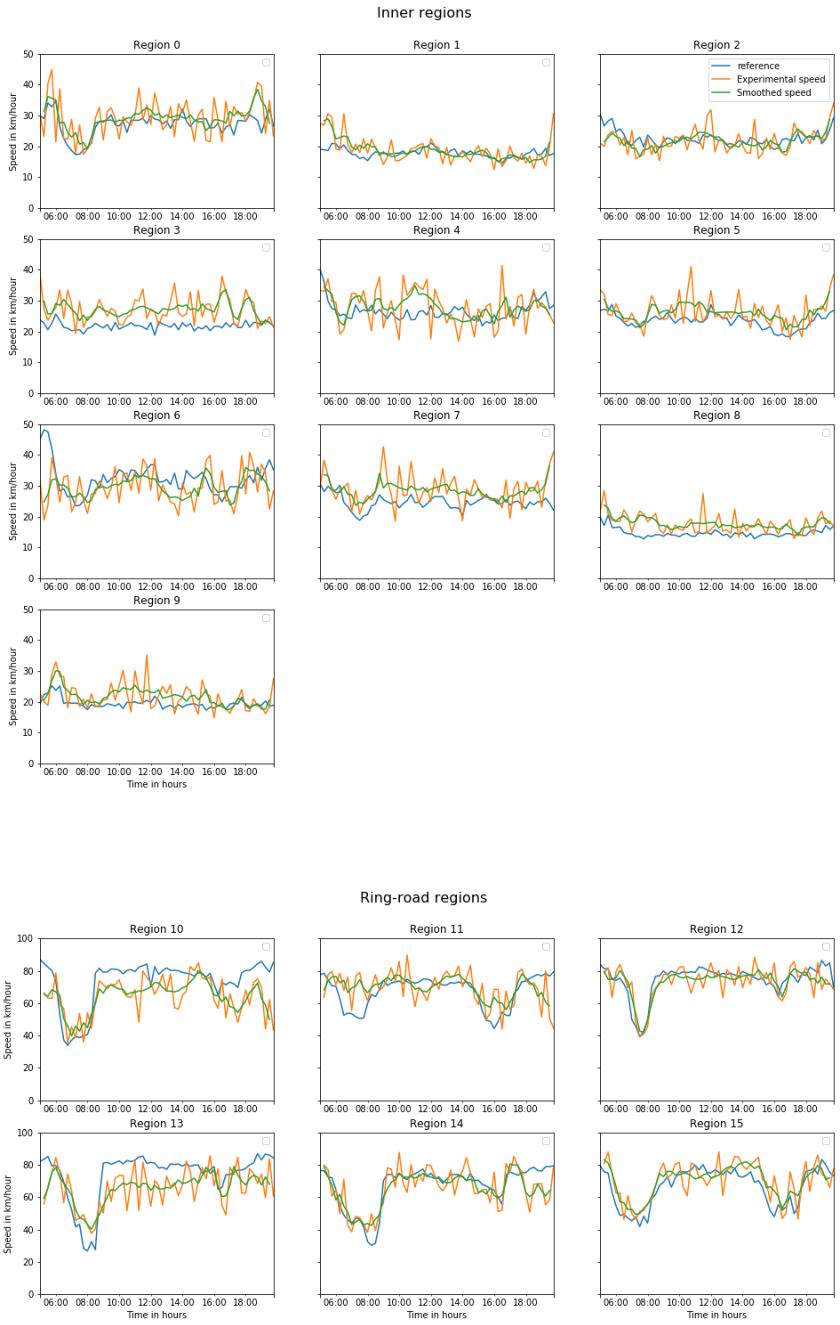



Figure 5.6: Speed estimation method applied to dataset $DS_0$ (trips with exact travel times))

correspond to the urban regions, while the last six ones characterize the speed dynamics on ring road.

For most of the inner regions, we observe that both the raw and smoothed speed trends well match our ground truth profiles, both during peak hours and off-peak periods. Regions 0, 1, 2, 5 and 6 display the most accurate speed estimations. Regions 3, 4 and 7 are those among urban regions where the raw results are the least stable, and for which the smoothing process is the least efficient. The traffic speed in these regions are slightly overestimated.

When it comes to the ring road regions (Region 10 to Region 15), we observe a more significant variability of the raw results, with saw-tooth raw speed estimations. For those regions, we observe that the speed estimates reproduce well the speed trends, generally following the ground-truth speed surges and drops during peak hours and matching the faster speeds in-between. Regions 12, 14 and 15 display the most accurate speed estimations. During peak hours, the raw results display important speed drops, and the increased sensitivity of the filter during assumed peak hours (6 AM to 9 AM and 3 PM to 7 PM) proves efficient to reproduce these dynamics while smoothing the results. In Regions 12, 14 and 15 the speed estimates reproduce well the speed drops. The speed estimates in Region 10 during the morning drop are also satisfactory. Region 11 is the most concerning at the peak time, as both its morning and afternoon speed trends are overestimated. After investigating this issue, we suppose that the divergence of the southern end of this section of the ring road, on the one hand, towards Region 12, and on the other hand, towards a freeway, with distinct road behaviors, could be the cause of this anomaly. In the future, we aim at looking more thoroughly at the characteristics of this region and explore how its network features might impact our results in this way. Although some smoothed speed trends in other regions miss reproducing the speed drops to their full magnitude (see in particular: Region 13, morning peak or Region 15, afternoon peak), the deviation from the baseline is much lower than the one related to Region 11, and the results remain satisfactory. As the raw speed estimates reach the lower speeds (Region 11, Region 15), modifying the filter during this time window to make it even more sensitive to lower speeds can be a way to reduce this gap and further improve results.

In-between the peak periods, the raw results follow the speed baseline and reproduce its speed dynamics. Regions 10 and 13 display the largest deviations from baseline with a general under-estimation of the speed during this time window. It is interesting to notice that those regions correspond to the opposite directions of the same section of the ring road: section North-East. Region 10 corresponds to the clockwise direction, while Region 13 corresponds to the counterclockwise direction. The reason for this under-evaluation of the speed is that both those regions are strongly connected to the north-western part of the ring road, which was not considered in this study as it mostly corresponds to tunnels. Hence, the process of filtering the scattered tracks related to this north-western section impacted the number of available trips in regions 10 and 13 more than the other ring road regions. In fact, we observe that the frequentation in those regions is, on average, 10% lower than in the other ring road regions This low frequentation leads to poor representativeness of the average travel time and generates a distance bias between actual and estimated travel distance per region and path. The other ring road regions display satisfactory results during this time window, for which the moving average succeeds in smoothing the raw results and their saw-tooth shape (Regions 11, 12, 14, and 15). However, this filter may be unsuitable if sudden and unexpected speed drops occur outside of peak periods. Despite this limitation, this filter was fast and easy to implement choice. In future work, we will explore other filtering techniques that both allow filtering the small saw-tooth instabilities of the results without neglecting the unexpected speed drop that may occur at any time of the day.

In Table 5.3, we detail the MAE and MAPE errors by period and region type for Day 1. Those errors are computed from the smoothed results. We observe that while the daily error is bigger in absolute value in the ring road regions than in urban regions, the absolute percentage error is smaller for the ring road. Generally speaking, the percentage errors are higher in peak hours than during the off-peak period. However, we also notice that its value is smaller for ring road regions than in inner ones, showing that the method is quite efficient in reproducing the fast-changing speeds of this particular kind of region.

Those results are interesting as they give a first insight into the potential of the method.

| | MAE (km/h) | | | MAPE (%) | | |
|---|---|---|---|---|---|---|
| | All regions | Inner regions | Ring road | All regions | Inner regions | Ring road |
| Full day | 4.727834 | 3.157841 | 7.344489 | 13.111325 | 14.274299 | 11.173036 |
| Off peak | 4.641103 | 3.090886 | 7.224797 | 12.040770 | 13.794607 | 9.117708 |
| Peak hours | 4.803374 | 3.216156 | 7.448737 | 14.043745 | 14.692095 | 12.963161 |
| Morning peak | 5.354016 | 3.665842 | 8.167639 | 16.491855 | 16.713291 | 16.122797 |
| Afternoon peak | 4.287147 | 2.794576 | 6.774766 | 11.748641 | 12.797223 | 10.001003 |

Table 5.3: Speed MAE and MAPE detailed by region and time window for Day 1

| | MAE (km/h) | RMSAE (km/h) | MAPE (%) | RMSAPE (%) |
|---|---|---|---|---|
| Day 1 | 3.780724 | 5.308405 | 10.711582 | 14.277904 |
| Day 2 | 3.698802 | 5.013979 | 10.599366 | 13.624562 |
| Day 3 | 3.706125 | 5.343630 | 10.516593 | 13.799747 |
| Day 4 | 4.061264 | 5.876281 | 12.163860 | 18.042898 |
| Day 5 | 4.078924 | 5.792345 | 11.312538 | 14.590598 |

Table 5.4: Daily speed errors when applying method to $DS_0$, using the actual trip lengths instead of static trip length estimates

Despite significantly lowering the information carried by individual trips (from GPS tracks to regional paths, and from exact arrival time to arrival period), the method reproduces the speed trends with limited errors. From the perspective of estimating traffic speed from temporally sparse data, this is a promising step.

However, we can identify several potential improvements. We already mentioned the improvements concerning the smoothing filter. The specific characteristics of Region 11 are also under investigation to understand how they impact the results. More generally, we can only stress the importance of the sample size. In fact, the number of individuals traveling along a regional path at each step must be large enough for the exogenously computed mean travel distance to represent the sample and for the sample's average travel time to represent the instantaneous dynamics along the path. When working with massive data, the amount of data available will ensure this representativeness and compensate for the low data information level. However, working with GPS data present the drawback of having to deal with a limited amount of tracks, and therefore, even more, a limited amount of tracks by regional path and period. This likely results in distance biases between the estimates and the actual average traveled distance, destabilizing the results. Hence, this case study can be considered a worst-case scenario in which the method requires us to work with limited access to trip information.

We explored the same speed estimation process from dataset $DS_0$ when replacing the static trip length estimates by the actual travel distances to validate those considerations. The speed trends for Day 1 are displayed in Figure 5.7, while the corresponding daily errors can be found in Table 5.4. The important improvement we observe, especially for Regions 10 and 13, confirms that a finer representativity of the trip length estimates should allow for more accurate results, thus limiting the gaps to the baseline. For this reason, and despite the limitations we mentioned, the method is very promising for an application to a way larger dataset.


Figure 5.7: Speed estimation method applied to dataset $DS_0$ (trips with exact travel times), using the dynamic trip lengths

| Avg IET | MAE (km/h) | RMSAE (km/h) | MAPE (%) | RMSAPE (%) |
|---|---|---|---|---|
| 4 | 7.894771 | 18.576767 | 10.773218 | 21.648403 |
| 8 | 12.970168 | 31.347989 | 16.233049 | 33.484246 |
| 12 | 16.918677 | 41.133133 | 20.600466 | 42.693440 |
| 16 | 19.920772 | 48.522253 | 24.008746 | 49.748655 |
| 20 | 22.204453 | 54.189049 | 26.608480 | 55.214964 |

Table 5.5: Speed errors on average over the week for each mean inter-event time selected as downsampling parameter

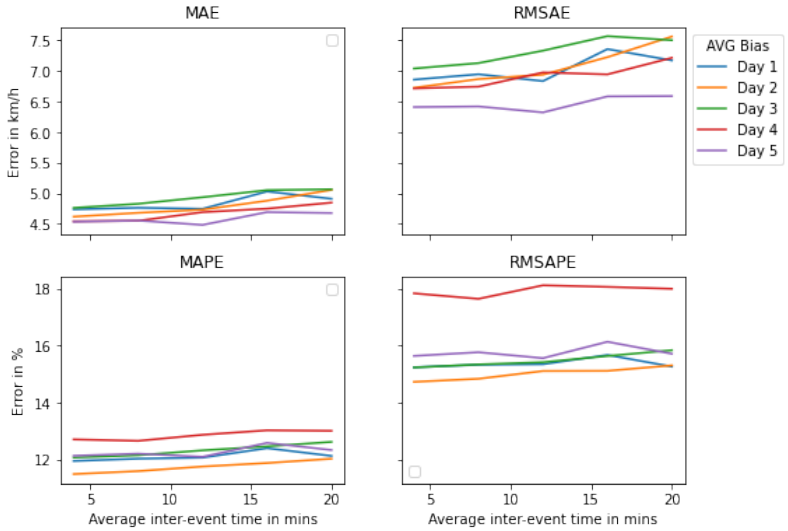



Figure 5.8: Evolution of daily errors with increase of average bias

### 5.4.2 Method application to trip data with biased travel time

In this section, we evaluate our method on the trip dataset $DS_1$ made of trips with biased travel times. First, we compute the zonal traffic speeds without applying the temporal bias removal. This allows evaluating the impact of the temporal imprecisions on the results. The variations of average errors over the week with average inter-event time are gathered in Table 5.5. We observe a significant increase in the errors, compared with Table 5.2. This shows how a bad estimation of travel times deteriorates the results' quality, even with a short average inter-event time and a limited travel time increase. It justifies the need for de-biasing in average the travel times. The following results are computed applying this de-biasing process.

We display in Figure 5.8 the evolution for each day of the different daily error indicators as a function of the average inter-event time. We observe that the error indicators are quite stable and rise slowly with the average inter-event time. On the contrary, when not expanding the data, the error increases quickly due to the increased dispersion of the bias distribution with the average inter-event time. This shows how important the sample size is and proves the capacity of a large dataset to compensate for the individual biases and imprecision and keep the bias removal process useful despite a large bias dispersion.

Figure 5.9 displays the smoothed results of the speed estimation for the five different average inter-event time values. The results for each value of average inter-event time almost fall into the same line, which confirms the aforementioned results. We observe that we are able, once again, to reproduce the traffic trends and dynamics.

In urban regions, the results are satisfactory in Regions 2, 5, 6, 8 and 9. In Regions 0,

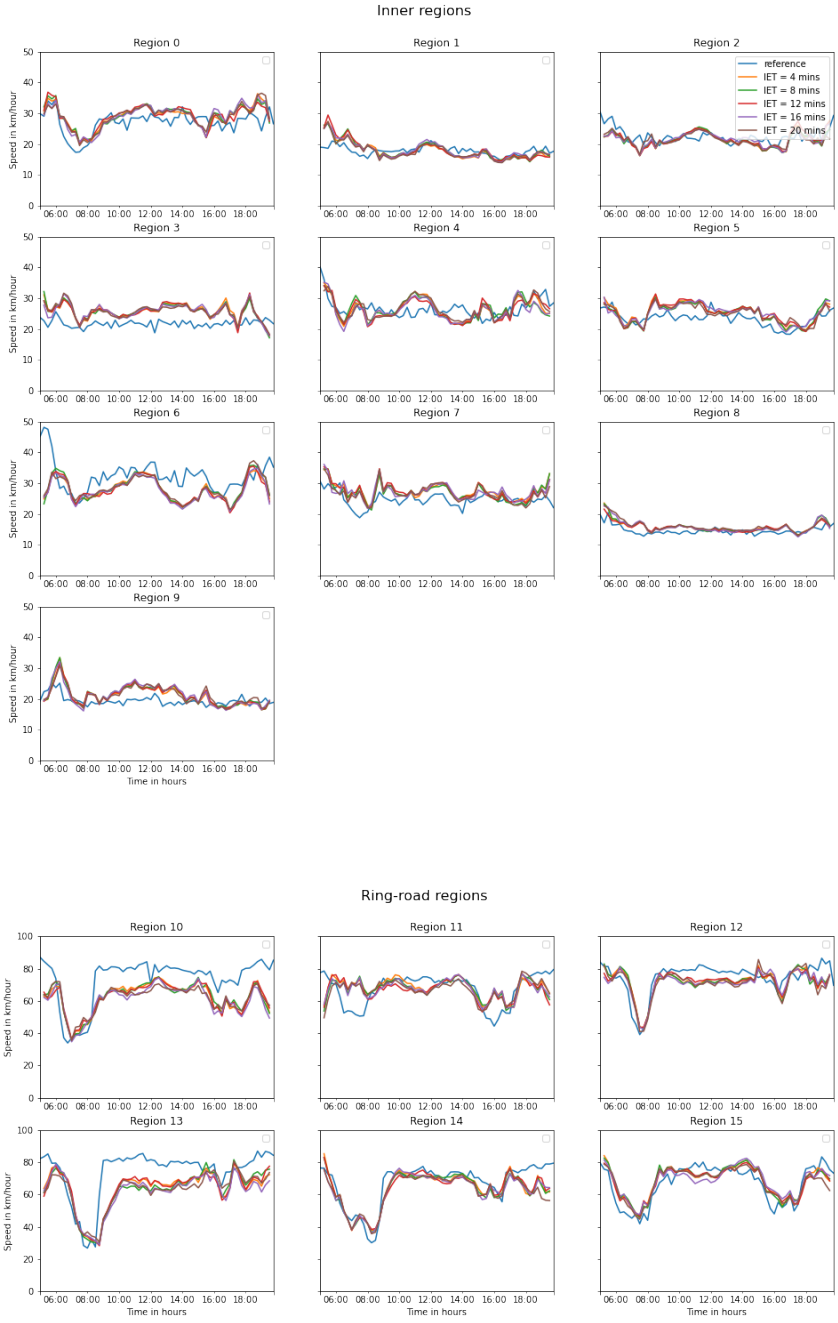



Figure 5.9: Speed estimation method applied to dataset $DS_1$ (downsampled trips) after average bias removal.

| | MAE (km/h) | RMSAE (km/h) | MAPE (%) | RMSAPE (%) |
|---|---|---|---|---|
| Day 1 | 4.911907 | 7.174379 | 12.124498 | 15.264945 |
| Day 2 | 5.057217 | 7.564531 | 12.024699 | 15.306312 |
| Day 3 | 5.067316 | 7.505059 | 12.617445 | 15.839542 |
| Day 4 | 4.848885 | 7.216612 | 13.006559 | 18.000039 |
| Day 5 | 4.678821 | 6.591109 | 12.331655 | 15.714796 |

Table 5.6: Daily speed errors when applying our method to $DS_1$ in the worst bias scenario (IET = 20 mins)

| | MAE (km/h) | | | MAPE (%) | | |
|---|---|---|---|---|---|---|
| | All regions | Inner regions | Ring road | All regions | Inner regions | Ring road |
| Full day | 4.911907 | 2.794057 | 8.441657 | 12.124498 | 12.113224 | 12.143288 |
| Off peak | 5.181909 | 2.790098 | 9.168261 | 11.769294 | 11.884125 | 11.577910 |
| Peak hours | 4.676744 | 2.797506 | 7.808808 | 12.433869 | 12.312762 | 12.635715 |
| Morning peak | 4.961260 | 3.236535 | 7.835803 | 14.194493 | 14.174446 | 14.227904 |
| Afternoon peak | 4.410010 | 2.385915 | 7.783501 | 10.783285 | 10.567434 | 11.143037 |

Table 5.7: Speed MAE and MAPE detailed by region and time window for Day 1 in the worst bias scenario (IET = 20mins)

1, 3, 4 and 7, the results are less consistent with the speed baseline. While the Region 3, 4 and 7 were already identified in the previous section as displaying less satisfactory results, the increases of the errors for the Regions 0 and 1 can be related to the introduction of the bias.

In the ring road regions, most estimated speed trends follow the speed baseline. The speed trends are particularly similar to the baseline in Regions 12, 14 and 15, although some speed drops are not reproduced with their full magnitude (Region 15, especially), but it was mostly already the case when working with unbiased data. Unsurprisingly, the results in Regions 10 and 13 remain underestimated in-between the peak periods, similarly to the case with unbiased data, but the speed drops are clearly observed. The estimation errors that we had already observed for Region 11 during peak hours in the case of unbiased data are increased when using biased data.

Finally, we further analyze the worst-case scenario results with an average inter-event time of 20 minutes. Table 5.6 displays the average errors observed for each day in this case, while Table 5.7 details the precise errors by region type and time window. Compared to the previous section results, we notice a general increase in the errors, although limited. The errors remain under a 20% limit when considering the daily RMSAPE, which is acceptable even though there is room for improvement here.

Overall, taking into account the errors previously introduced by upscaling of the GPS tracks to the regional path, working with biased trips seems to have a limited negative impact on the result. Despite the low quality of the trip information at this stage, the results are very encouraging. Therefore, the room for improvement includes the reduction of errors at each stage of the process. This ranges from the representativeness of trip length and time estimates, to the filtering process, to a more refined understanding of the impact of internal speed dynamics in the results.

| Avg IET | MAE (km/h) | RMSAE (km/h) | MAPE (%) | RMSAPE (%) |
|---|---|---|---|---|
| 4 | 6.402842 | 14.587325 | 9.344040 | 17.957539 |
| 8 | 10.950956 | 24.768931 | 14.947251 | 28.044663 |
| 12 | 14.044534 | 32.729041 | 18.393624 | 35.744618 |
| 16 | 16.329243 | 38.268281 | 21.118529 | 41.143766 |
| 20 | 17.901683 | 42.022657 | 23.004366 | 44.958863 |

Table 5.8: Speed errors on average over the week for each mean inter-event time selected as downsampling parameter

### 5.4.3 Method application to trip data with both biased arrival and travel time

In the preceding section, we have analyzed our results when using trips with a biased travel time information. However, UADP data not only display biases in the travel time, but on the arrival time as well. In this section, we therefore consider this additional bias on the trips by exploiting the $DS_2$ dataset, and explore the impact of the methods we propose on such results.

First, we compute the results of our method on dataset $DS_2$ when handling the travel time bias only. The results are displayed in Table 5.8. Compared to Table 5.6 for instance, which represents the average errors we obtained for each day in the worst case scenario, those results display a new significant increase of the errors. Since the arrival times are de-skewed, this increasing of the errors is related to the arrival time bias only, which results in mixing together users traveling at different periods and in erroneous travel time estimations. Without surprise, we can observe that the larger the average inter-event time is, the larger the errors are, because trips are shifted further away from their actual travel time period.

Therefore, handling this arrival time uncertainty seems necessary, as it was previously done with the travel time. This is what we address in the second part of this section. Figure 5.10 displays the results obtained once we shift back each trip's arrival time by $\mu_X/2$, remove users with bias larger than twice the average bias and filter regional paths that represent less than 30 individuals.

Although we still observe a sensibility to the average inter-event time (and average bias), the results are contained within much lower bounds than the ones observed in Table 5.8, showing the filters' efficiency in limiting the arrival time bias impact on the results. However, compared to Figure 5.5, we observe a larger increasing of the error with the average bias, which can be explained by the fact that the larger the average bias is, the larger the variance, resulting in an increased data shuffling.

In Figure 5.11, we display the speed estimation results obtained for each inter-event time value on Day 1. These plot display results that can reasonably be compared to the ones exposed in 5.9. Table 5.9 precise the daily results in the worst case scenario, while Table 5.10 precise the errors by region and time period. Overall, the increasing of the errors compared to Tables 5.6 and 5.7 is limited, which confirms the viability of our method for estimating regional traffic speeds despite low-quality path information and fully biased temporal features. In particular, these latest analyses demonstrate the utility of implementing filters at the individual and equation levels to compensate for the temporal biases of the data. This suggests that these filters will have great potential when it comes to handling large amounts of data, which we are eager to verify.

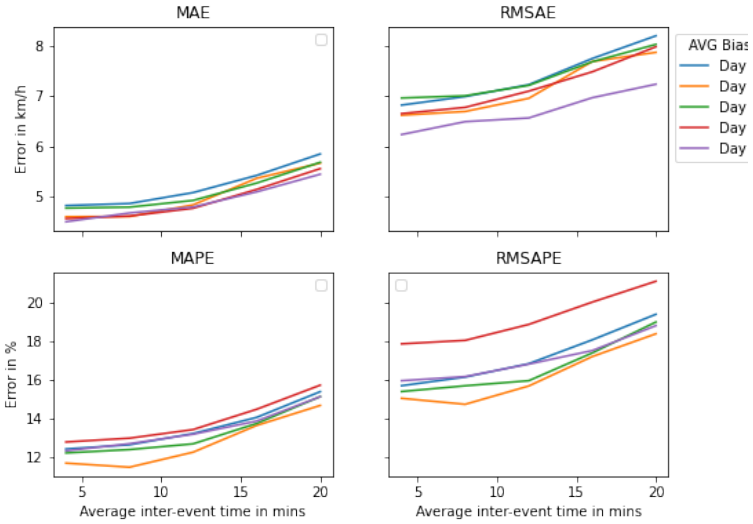



Figure 5.10: Evolution of daily errors with increase of average bias

| | MAE (km/h) | RMSAE (km/h) | MAPE (%) | RMSAPE (%) |
|---|---|---|---|---|
| Day 1 | 5.843820 | 8.214792 | 15.377055 | 19.398781 |
| Day 2 | 5.659251 | 7.883027 | 14.660121 | 18.385150 |
| Day 3 | 5.676346 | 8.042682 | 15.119238 | 18.995601 |
| Day 4 | 5.549316 | 7.994296 | 15.714570 | 21.116936 |
| Day 5 | 5.436428 | 7.243572 | 15.116918 | 18.807797 |

Table 5.9: Daily speed errors when applying method to $DS_2$, when in the worst bias scenario (IET = 20 mins)

| | MAE (km/h) | | | MAPE (%) | | |
|---|---|---|---|---|---|---|
| | All regions | Inner regions | Ring road | All regions | Inner regions | Ring road |
| Full day | 5.843820 | 3.720738 | 9.382290 | 15.377055 | 16.573524 | 13.382939 |
| Off peak | 5.999818 | 3.771147 | 9.714271 | 14.918967 | 16.531425 | 12.231538 |
| Peak hours | 5.707951 | 3.676834 | 9.093144 | 15.776034 | 16.610191 | 14.385773 |
| Morning peak | 5.621608 | 4.077899 | 8.194454 | 17.050427 | 18.541036 | 14.566079 |
| Afternoon peak | 5.788897 | 3.300836 | 9.935666 | 14.581290 | 14.800023 | 14.216735 |

Table 5.10: Speed MAE and MAPE detailed by region and time window for Day 1, when in the worst bias scenario (IET = 20mins)

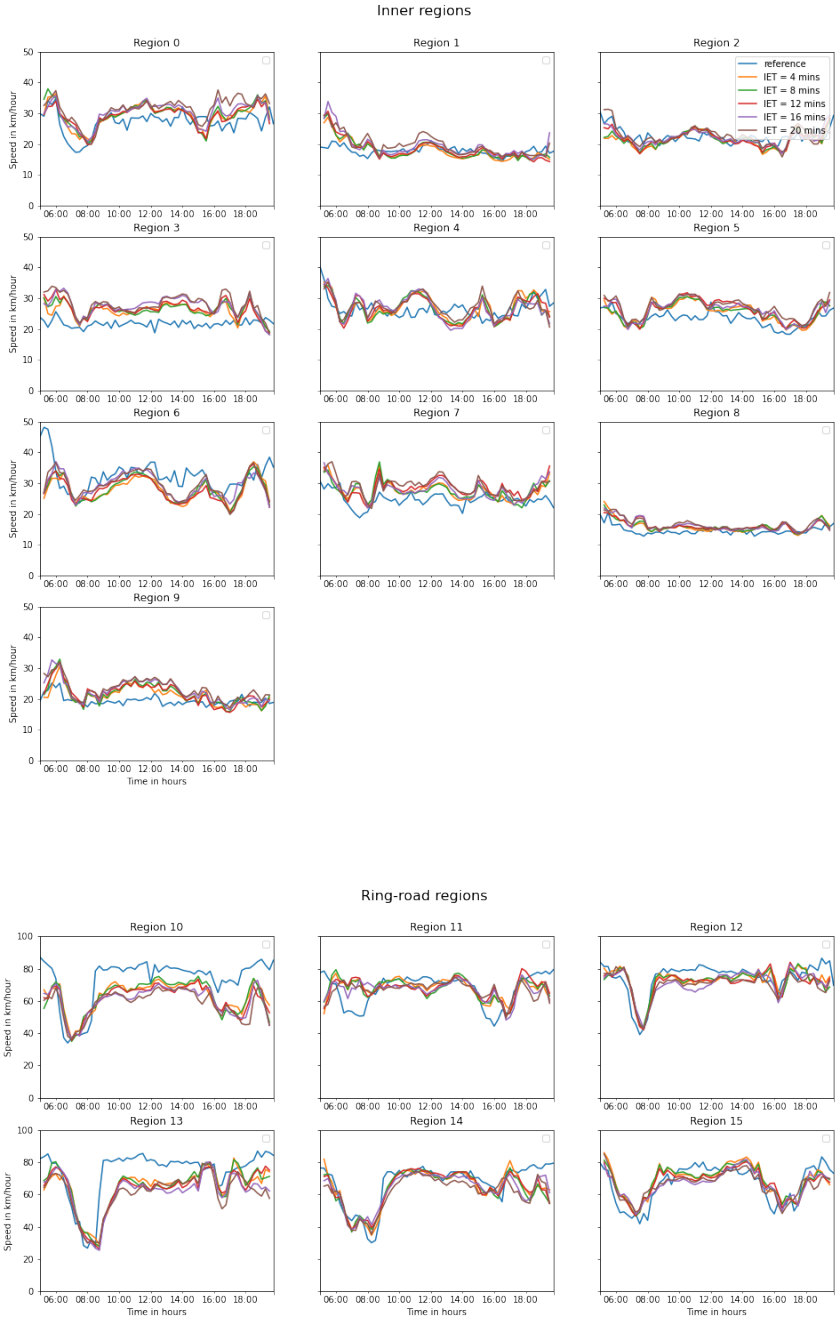



Figure 5.11: Speed estimation method applied to dataset $DS_2$ (fully biased trips) after average bias removal and arrival time correction.

## 5.5 Conclusion and discussion

This chapter has proposed a new methodology for estimating the dynamics of regional traffic speeds from user-activity-dependent positioning data. The trips extracted from these data present the challenging issue of being temporally biased, making the individual traffic speed difficult to estimate. To address this issue, the method we propose first relies on the definition of a proper data resolution scale, both on the temporal and spatial dimension, which is used to group and aggregate the user-activity-dependent positioning data. It especially requires the partitioning of the studied area in sub-regions characterized by homogeneous traffic. Such partitioning allows defining basic trip features, such as regional paths, which allow the identification and aggregation of similar trips. This aggregation allows a systematic, exhaustive, and robust estimation of average travel times throughout the network and at each time step through the fusion and de-skewing of individual travel times. Finally, provided that estimates of the regional trip lengths have been performed beforehand, the travel time estimations are jointly analyzed to deduce the underlying regional traffic speeds. This structure of the method is particularly fitted to any massive but temporally sparse data input, as it requires very little temporal or itinerary information at the individual level and considers the inherent temporal bias that characterizes trips extracted from those data.

Applying this approach to downsampled GPS data offers a controlled environment to evaluate the different degrading steps of our approach. First, despite reducing the available GPS trips to minimal temporal and path information, the method could reproduce the speed trends throughout the day, especially the fast-changing dynamics observed in the ring road regions. The moving average smoothing filter that we implemented in this article was proved to be efficient to smooth the period-to-period instabilities of the results. More elaborated filters can replace this one in future work. Despite these satisfactory results, two regions, in particular, displayed underestimated speeds, which we related both to their representation level and inner dynamics. After the introduction of individual temporal biases on the travel time, we repeated the experiment. Different bias models were explored by making the average inter-event bias vary between 4 (best case) to 20 (worst case) minutes. We showed that, provided that the amount of data was sufficient for the mean bias to be representative of the individual sample, the system could be de-biased and return satisfactory results, although we noted that the error slightly increased. In the last step, the method was applied on trips for which both the observed travel and arrival times were biased. At each step, we have identified methodological options that could help to reduce these errors. Working with a large amount of data was identified as an essential requirement of the method. Indeed, it ensures both good reliability of the systems' equations and a correct de-biasing process, especially when working with datasets characterized by long average inter-event time. The sample size was an issue in our case study, in which we had to deal with both low data availability (related to the GPS data source) and the poor quality we imposed on the data to replicate the characteristics of the data UADP. This problem was circumvented by artificially increasing the size of the trip data set by duplicating each displacement 100 times. When working with massive UADP datasets, this problem should no longer arise because the amount of data per regional path will be much more significant, allowing for greater representativeness of displacement lengths and adequate management of bias dispersion. To further investigate the first aspect, we also showed how more accurate dynamic trip length estimates could reduce errors. It is a promising research direction as several studies in the literature have shown that regional trip lengths are relatively stable from day to day but can experience variations within days related to the congestion spreading [Batista et al., 2021a, Paipuri et al., 2021].

In future works, we first would like to explore the sensibility of our method to the

different parameters such as the size of the regions or the period duration. We also look forward to testing the robustness and the portability of our method in other geographical contexts. Above all, we would like to address the critical assumptions concerning the travel time bias we made in this study. The average value was assumed to be known and considered static in time and invariant to space. While the temporal characterization of inter-event times in UADP data has been explored in several works, the specific question of the bias existing between observed and actual travel times when working with trips derived from these data has, to the best of our knowledge, never been explored by the literature. Therefore, the assumed characteristics of the temporal bias are difficult to confirm or invalidate. The leads for the estimation of such a bias are also limited. Although our method only requires an estimate of the average bias and not a full characterization of its distribution, the lack of literature on the subject limits the immediate application of this approach. In this chapter, we have proposed a simplistic model relating this bias with the inter-event time distribution. Although the objective of this model was mainly to provide a methodological context for the sub-sampling of data, we believe that the characterization of this bias does indeed require relating it to the inter-event time distribution. We would like to investigate this question further.

# Chapter 6

# Towards emission calculation from CDR data: a global framework

## 6.1 Introduction

For this thesis, we have defined two objectives. The first objective was to develop methods for estimating the traffic variables required to estimate traffic-related pollutant emissions. Our second objective is to integrate these methods in a global framework allowing the estimation of these emissions. It is the purpose of this chapter.

We discussed the need to select the traffic variables according to the targeted emission model implemented at the end of the modeling chain. We identified specific emission models (macroscopic models) that are adapted to the data characteristics, the urban scale of emission assessment, and the decision-making context in which the tool is intended to operate. In these emission models, the estimation of the emission factors relies on average traffic speeds and vehicles features, such as motorization or emission standards. These emission factors characterize the emissions produced per unit of distance traveled. Therefore, the estimation of global emissions, on a regional and urban scale, requires the multiplication of these factors by the total traffic volume. For the record, the emissions of a pollutant $k$ are estimated with:

$$E^k = TTD \cdot F^k(V) \tag{6.1}$$

where $TTD$ is the Total Travel Distance (in km), $V$ for the traffic speed (in $\mathrm{km\,h^{-1}}$) and $F^k(V)$ the emission factor (in $\mathrm{g\,km^{-1}}$) (cf. Introduction).

It relates the emissions with the traffic speed $V_r$ and the traffic volume $TTD_r$. So far, this thesis has focused on proposing methods to estimate these two variables considering the full spatial extent and all population categories, and overcome the limitations of state-of-the-art. In Chapter 1, we proposed a method for partitioning the communication network to build zones of minimal size and population. In Chapter 2, we focused on characterizing the population and the urban network. On the one hand, we categorized the individuals according to different presence profiles in the area. On the other hand, we identified most popular regional paths and characterized their distances. Part II of the thesis focused on reconstructing the population mobility and estimating the traffic volume. Chapter 3 focuses on the reconstruction of individual and regular mobility, while Chapter 4 proposes a more aggregated method of traffic volume estimation for an irregular population. Finally, Chapter 5, the first chapter of the last part dealing with emission reconstruction, proposed a method for estimating average speeds, an essential variable for estimating emission factors in aggregated models. In this chapter, we propose to relate

and synthesize these methods into a global modeling chain. This chapter proposes a global vision on this thesis contributions through the Introduction of the general framework.

This chapter is organized as follows. Section 6.2 presents the framework and describes how unit methods connect with each other. Section 6.4 presents a number of perspectives for the continuous improvement of this framework that we consider to be priorities. Section 6.5 concludes this chapter.

## 6.2 Global framework

We propose a global modeling chain, which takes as inputs CDR data, road network data, and census and survey data. It results in the traffic variables required for emission estimation, the traffic speeds and the traffic volume. It is organized around the speed estimation process, and the parallel treatment of two distinct categories of users with different approaches. This distinction makes it possible to considerably broaden the scope of the users considered and to jointly process diverse mobility profiles. This global modeling chain is illustrated in Figure 6.1, where oval boxes correspond to the data flow, while square boxes correspond to data processing steps. The bold boxes indicates the processes based on historical analysis. In red, we represent the interfacing of our framework with the steps related to the detection of the transport mode and the estimation of the temporal bias of the data. These elements are necessary for the effective estimation of emissions, and are discussed in the next section. This general framework falls roughly into six majors steps, represented with light gray blocks in the figure, which mostly correspond to the different chapters of this thesis:

1. Categorizing individuals according to their observed presence profiles based on longitudinal data (observations over several months) and complementing this categorization with a scaling method adapted to each user class.

2. Analyzing the road network to identify dominant routes, and related regional lengths.

3. Reconstructing the individual mobility of regular users based on this analysis and the previous categorization. This reconstruction focuses on two aspects of their mobility: on one hand, the enrichment of activity chains based on the identification of historical mobility routines; on the other hand, the reconstruction of travel trajectories on the other.

4. Reconstructing mobility for non-regular users (non-regular locals and tourists). This second approach allows direct estimation of the distances traveled by the sample, without the need to analyze more detailed mobility patterns.

5. Estimating traffic speeds by fusing large amounts of temporally-biased trips. Trips should be separated by transportation mode, or even by vehicle type, in order to estimate accurate, mode-specific speeds, and adapted emission factors. This specific issue was out of the scope of this thesis.

6. Estimating the emissions factors based on the traffic speed and relate them to the total travel distances for a global estimation of the emissions.

These steps are related as follows.

**Step 1**: The initial CDR user categorization allows to distinguish users between regular and non regular travelers. It is based on a longitudinal analysis of the CDR data, and calibrated resorting to macroscopic indicators derived from local surveys and census data. It results in a segregation of users according to presence profiles (residents, commuters and

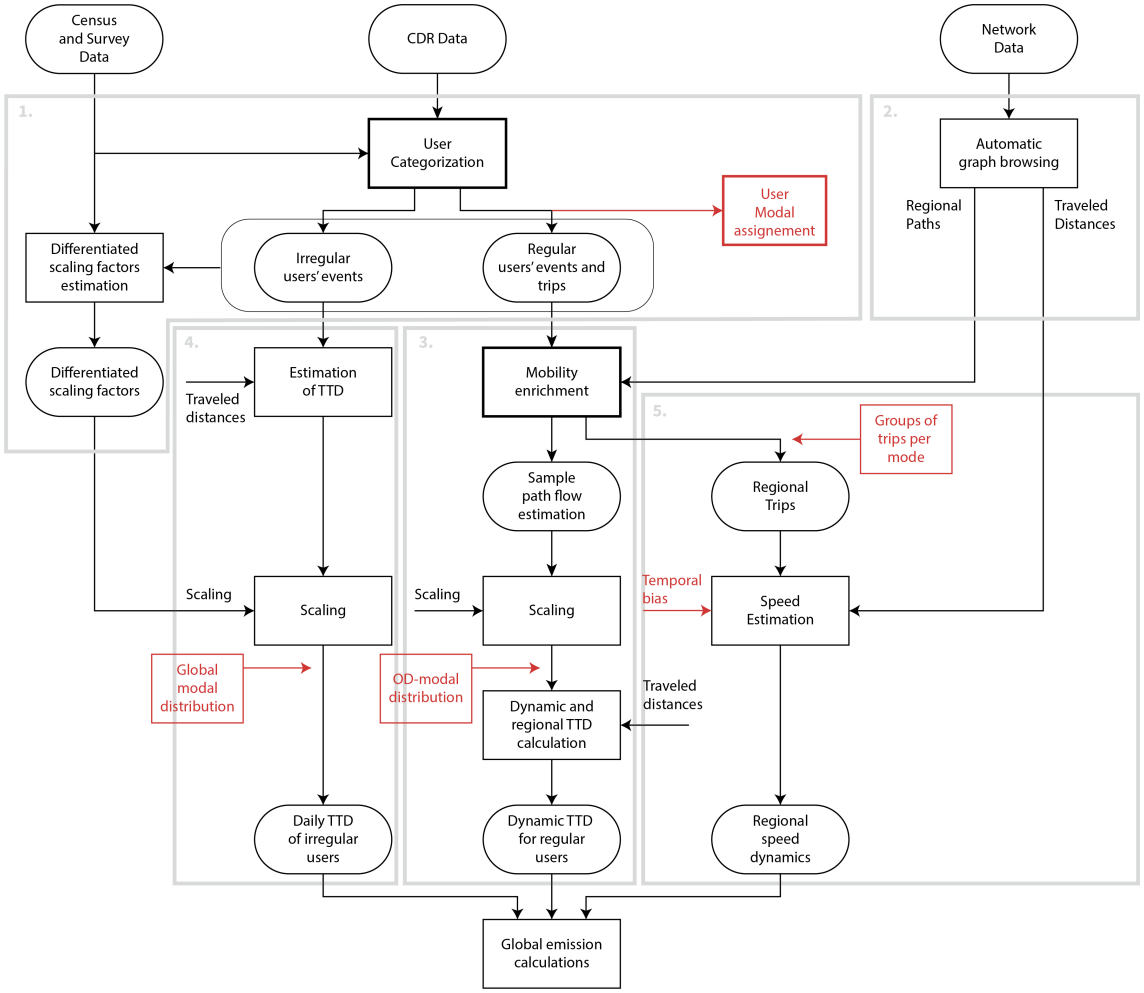



Figure 6.1: General framework proposition

locals) and mobility regularity. In this step, the formalization of scaling factors definitions for the different mobility profiles provide a way to upscale the mobility characteristics of the users considered. These categories can be gathered into two main user classes, regular and irregular users. This macro-classification supports the development of two different mobility reconstruction approaches, one for each.

**Step 2**: Before moving on to those, another fundamental step supports the rest of the framework (Figure 6.1, step 2). It is the systematic analysis of the road network, based on a state-of-the-art approach [Batista et al., 2019] that rely on road network data only. The two low-level modifications of this method that we propose allow for scaling up and applying it to the large spatial dimensions of our case study. It results in the identification, for each regional origin-destination couple, of the prevailing regional paths between them, and in the estimation of the average regional trip lengths along those paths.

The reconstruction of the distances traveled by users is based on two parallel processes, designed to treat regular and irregular users separately.

**Step 3**: In the first case, the framework uses an individual reconstruction of mobility. At a first level, this reconstruction consists in enriching the daily activity chains on the basis of a historical analysis. At a second level, it consists in assigning regional trajectories to the observed trips, based on the set of pre-identified prevailing paths and the observations of the CDR data. Expanding the results with the scaling factors provided by step 1, and multiplying the regional flows by the average distance provided by step 4, allows to estimate the total regional distances traveled by regular users.

**Step 4**: For irregular users, the proposed approach is more collective and spatially aggregated. The aim is to focus on a small sample of users in order to limit the impact of time bias on the evaluation. The data are computed from the activity chains and the calibration of a detour ratio function. After applying the scaling factors provided by the Step 1, it results in a macroscopic estimation of the total travel distances.

**Step 5**: The last step of the framework that was developed in this thesis consists in estimating the average traffic speeds. It relies on the integration of trip data extracted from the mobility analysis, on an estimate of the average time bias of the monitored population, and on the regional trip lengths resulting from Step 2. It results in regional speed dynamics.

**Step 6**: The interfacing of this framework with a method of detection of the modes of transport will allow, on the basis of the variables in outputs of the Steps 3, 4 and 5, the calculation of the emissions relating to the traffic of the zone and the population studied.

Although the issue of modal identification was out of the scope of this thesis, we discuss the perspectives of its implementation and its connection to the proposed framework in Section 6.4.2. This section also discusses the estimation of the time bias.

Before that, the next section lists the contributions of these different steps, and their integration into this global framework.

## 6.3 Contributions to the research objectives

This section related the contribution of this framework and of the method developed throughout this thesis to the research objectives listed in Introduction.

- In contrast to many works in the literature based on cell phone data that strive to relate the data to the road network, we propose a reconstruction method based on an intermediate regional scale on the one hand (regular users), and collective motives on the other hand (non-regular users). Our entire approach is guided by this decision. In particular, it guides one of the first analyses of this thesis, which focuses on characterizing the routes taken at a regional scale on a network, and the average

trip lengths these routes. In this sense, we adapt a method from the literature to the broad geographical dimensions of our study.

- In order to estimate global and exhaustive total distances traveled, we propose to base our mobility analysis on a categorization of individuals. The type of treatment subsequently applied to users depends on the category of the population to which they are assigned. Articulating such a classification method with different mobility analyses is an original proposition of this framework and, which allows to enlarge the scope of the users profiles considered in the analysis, while the CDR-based literature tends to focus on users who are residents of the studied area and regular users. In addition to the actual participation of these travelers in traffic volumes and thus emissions, they may also be the source of specific mobility patterns, of which resident individuals are not representative. We complement this categorization with the calculation of scaling factors adapted to each category considered.

- The temporal irregularity of cell phone data suggests an enrichment of this mobility if one wishes to estimate exhaustive and non-biased traffic volumes. This question is not often raised in the literature, and in particular in works that focus on the estimation of OD matrices, but it is fundamental because fragmented data chains result in erroneous calculations of OD flows. As a consequence of the diversity of the users we include in the approach, we propose two approaches to reconstruct this mobility, with different granularity of analysis.

- While many works in the literature focus on OD matrix estimation without taking into account the fragmented nature of the data, we first propose a heuristic for enriching incomplete activity chains for regular users.

- This approach is then completed by a method of enriching the paths taken, based on a partial map-matching of the users and the estimation of path-flow distribution. The choice of a regional approach is a strategic element of this step.

- We propose a more collective and spatially aggregated approach for non-regular individuals. It consists of estimating total distances traveled directly from the users' activity chains, taking advantage of the concept of detour ratio.

- We develop an average speed estimation method for temporally biased positioning data, that relies on the fusion of large amounts of biased trips and on an estimation of the average bias. We test the method on a set of artificially-generated temporally-biased trips and obtain satisfactory results, which bodes well for an implementation on real CDR data.

- Finally, a complete articulation of these different contributions was proposed in the previous section. To the best of our knowledge, it is the most integrated methodological framework for estimating the traffic variables required for emission calculations. It especially provides numerous keys to handle the data sparsity and limits its impact on the variable estimations.

## 6.4 Perspectives

In each previous chapter, we have highlighted the limitations of our developments and discussed related future work that we would like to pursue. This section rather intends to focus on the further development of the framework itself, discussing the remaining gaps that need to be filled to render a fully automated and integrated process for emission calculations.

### 6.4.1 Temporal bias analysis

The speed estimation method that we propose in the chapter is based on a fusion of trips extracted from the CDR data, and which are thus biased temporally. It requires to de-bias them on average, and thus to estimate the average bias of the studied trips. This input is represented in red in Figure 6.1. This issue has not been addressed in this thesis, yet, for a reliable estimation of the speeds, such a study would be necessary. Here we discuss two possible approaches.

The first one requires the availability of ground truth data, and in particular floating car data (LBNS). In theory, these data would allow to identify with precision the trajectories of individuals, and thus the moments of departure from the origin and arrival at the destination. However, the implementation of such validation processes for CDR data analysis is complex. It often requires the development of suitable mobile applications, the distribution of these applications to users, the agreement of these users to cross-reference their GPS data with their cell phone data, the specific sharing of these users' data by cell phone operators. In the end, these validation methods often result in small sample sizes. While they may allow the validation of mobility reconstruction methods, the sample size would certainly be a limitation in identifying laws governing temporal biases between observed and actual mobility.

Another approach could be explored, based on the advanced analysis of the CDR data itself, and in particular of the days corresponding to the same activity chain. Focusing on the most temporally regular activity chains, Although we focused on the spatial dimension of the regularity of activity chains, a significant number of individuals should exhibit activity chains with low temporal variability. The systematic comparison of the temporal activity chains of these could allow to specify the latent usual departure and arrival times. Comparing them to the times observed from the raw data would provide an estimation of the temporal bias of the trips extracted from these data. This will be the subject of future studies.

### 6.4.2 Integration of mode detection methods

Estimating features related to the transportation mode, such as individual trip modes or collective modal distribution, was kept out of the scope of this thesis. Specific researches have been lead on this subject [Huang et al., 2019], and, in the context of the Green City Big Data research and development project, this question is currently being addressed by a partner of the project. However, towards a global and integrated traffic-related emission estimation framework, plugging a mode estimation method will be required. As the emission factors are broken down by transport mode and vehicle class, it is desirable to estimate specific traffic variables for each mode. This implies an identification of the transport modes at different granularity. Regarding the estimation of traffic volumes, collective information, characterizing the distribution of modal choice by path, origin-destination pair, or over the whole study area, may be sufficient to distribute distances by transport mode. For the estimation of speeds, it is also desirable to distinguish trips by mode, considering, for example, that the regional speed is not the same depending on whether the user is traveling by car, motorcycle, or on exclusive right-of-way public transit. However, a systematic, on-line estimation of transport modes per trip would likely be redundant and costly, and probably often unsuccessful considering the significant time bias in the data. Instead, our suggestion would be to identify, on the basis of the historical approach described in the previous section, the main mode(s) for each studied individual and to consider this information as stable when processing new days of data. Indeed, the literature has shown that the stability of travel behavior extends to modal choice. In the long term, the implementation of this method may involve integrating the characterization

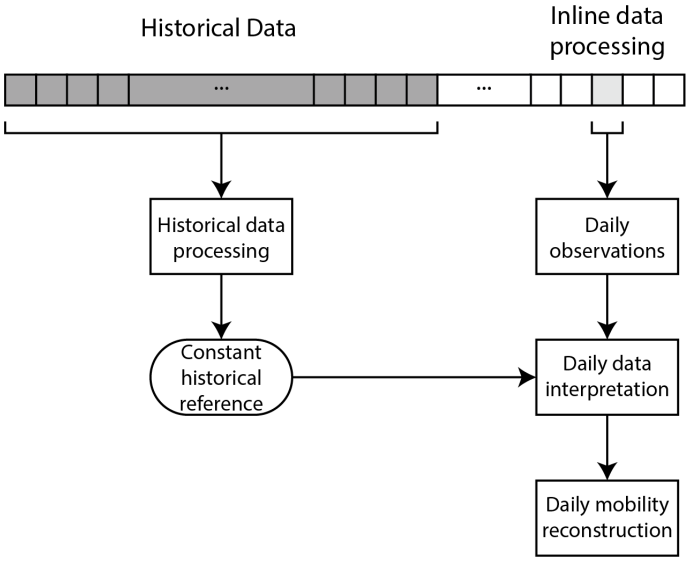



Figure 6.2: Historical data processing for day-by-day mobility reconstruction

of the transport mode into the analysis of the routines. However, at this stage of research, it seems reasonable to assume that the different motorized modes have the same regional average speed dynamics. This assumption will support a first estimation of the traffic-related emissions, before being reconsidered.

### 6.4.3 On field implementation: periodic learning of historical patterns

Two of the unit methods developed in this thesis rely on individual features that were learned from a longitudinal analysis of historical data. First, the categorization of users requires analyzing their data history to determine whether they are visitors, commuters or residents, and in the latter two cases, whether their mobility is regular or not. Second, the enrichment of the regular users' mobility relies on identifying complete and recurrent activity chains grouped into mobility routines, whose probability is estimated. In Figure 6.1, we have highlighted the related processes with bold boxes.

In both cases, learning on the historical database provides individual user characteristics that allow the processing of new daily data (categorization, orientation towards the adapted mobility reconstruction approach, enrichment of activity chains). Up to this point in the implementation of this framework, we have considered that the features extracted from this historical analysis were stable over time. For instance, the users' presence and mobility profiles are considered constant. The principle of this process is illustrated in Figure 6.2. However, from the perspective of implementing this framework for practical, long-term mobility studies, this use of a learning phase raises several questions. In particular, it raises questions about the long-term validity of conclusions drawn from the learning phase, as well as the need for and possibility of continued learning.

On the one hand, we illustrated Chapter 2 the sensitivity of the categorization of individuals to historical data availability. Due to the unequal amount of data between these individuals, different lengths for the learning phase results in different classifications of the users. The longer the historical period considered, the more accurate the resulting classification. It is very likely that the construction of the mobility profiles suffers from the same issue. Because of lower communication rates, some users might require more time for their mobility routines to be identified. Therefore, it seems relevant to seek to regularly enrich the data history. It would allow to update the users' various features in light of the new data, and to progressively improve the quality of the resulting estimates.

In addition to the issue of unevenness in the amount of data between users, it is obvious that the mobility characteristics, at an individual or collective level, can change over time. At an individual level, residence or work change can impact the mobility and make the historical analysis (presence or mobility profile) obsolete. On a collective scale, major events can cause influxes or desertions of mobility (celebrations, crises, ...). The recent Covid-19 crisis is a very good illustration. Therefore, in addition to a regular update of historical references, a monitoring of individual and collective mobility deviations and a flexibility in the application of a historical reference to current observations would certainly be required when using this framework in practice.

The implementation of this last suggestion obviously requires the construction of numerical indicators to measure the predictability of daily data by historical reference. At an individual scale, this can take the form of measures of intersection of daily mobility with historical routines, or mobility metrics such as radius of gyration. At a collective scale, the stability of origin-destination matrices and the total distance traveled before historical completion look relevant to characterize regularity. Their historical monitoring may allow the detection of mobility anomalies. The literature is rich in methods specific to the detection of crowd events and may inspire the further development of this framework.

In the absence of major detection of mobility changes, two approaches are possible regarding the regular update of the historical patterns. The first one consists in proceeding to a new learning phase at regular intervals and to draw new mobility references from it. This approach is probably the least expensive. A specific study will be able to evaluate the most suitable time period during which the previous history remains mostly valid. Another approach would be to update individual histories for a local part of the population if the monitored mobility indicators differ excessively from the historical reference indicators.

## 6.5 Conclusion

This chapter aimed at presenting a global framework integrating the different methods developed in the thesis, and connecting them in order to estimate the traffic variables necessary for the estimation of the related atmospheric emissions. The proposed framework uses three types of data (CDR, road network, census and survey data) to estimate average regional trip length, total regional and citywide trip distances, and traffic speeds. The initial methods for characterizing the network and the sample support the rest of the data processing chain. The following steps are organized around a new method for estimating traffic speeds and two concurrent approaches to mobility reconstruction, with distinct and complementary levels of analysis.

Before a complete implementation of this framework on the network of Cali, it still needs to be enriched by a characterization of the temporal biases induced by the data in the mobility extraction, and by methods of transport mode detection. Then, towards an on field application of the method, we suggest adapting the workflow for periodic or continuous learning of the mobility pattern in order to process in an automated way a variable and sometimes unpredictable incoming mobility.

# Conclusion

This thesis focuses on developing a global methodological framework to support the estimation of traffic-related air emissions in urban areas from a specific type of mobile phone data, CDR data. These data have the advantage of being massive, available, and with a structure adapted to the study of individual mobility, which explains why they have been the subject of numerous studies over the last decade. Because of these qualities, they constitute an alternative data source to the classical methods of mobility analysis (mobility surveys or loop detectors, for example).

However, their temporal dispersion and low spatial resolution pose several problems when estimating specific traffic variables. First, spatial resolution is a limitation to characterizing fine trajectories and travel distances. Second, the irregular sampling rates are a significant limitation to evaluating global OD and path flows. Third, these characteristics prevent the evaluation of accurate travel times and the estimation of individual traffic speeds. Finally, the lack of additional statistics on the study population limits the interpretation of its mobility and often forces the analysis to be restricted to specific user profiles, like residents.

The framework elaborated in this thesis gathers methods derived from the literature (and adapted to our problem) and original solutions into a consistent modeling and data processing chain. The details of the contributions of this thesis are presented in the previous chapter. We list them here briefly. They include:

- formalizing up-scaling methods adapted to different user profiles for their integration in mobility analyses;
- adapting an automatic graph browsing method for the characterization of a road network, its prevailing paths and its regional average travel distances;
- developing a path flow estimation method based on a coupled activitiy-chain/trajectory reconstruction approach;
- designing an alternative approach for estimating the total distances traveled on the network via the concept of detour ratio;
- developing an innovative approach for estimating the traffic speeds.

The framework is in itself an essential contribution of this thesis since it provides the keys for estimating air emissions from CDR data and limited complementary data.

The perspectives of the framework are established at two levels, at the unit method level, and at the overall framework level. In the course of the manuscript, we have identified for each of them perspectives of improvement and research. According to us, the main ones are the following.

- **User categorization** (Chapter 2): Improving the classification of individuals by taking into account wider range of mobility behaviors and implementing more flexible user categorization method will likely provide more robust user categories;

- **Individual mobility reconstruction** (Chapter 3): Refining the methods for enriching the users' activity chains will allow to better take into account the users' specific mobility and communication habits. Moreover, designing mobility routines that differentiate between weekdays and weekend appears as an essential step towards a more flexible reconstruction of the daily individual mobility patterns.

- **Collective travel distance reconstruction** (Chapter 4): Coming back on the assumption that most active irregular users are representative of the remaining irregular users from the daily travel distance perspective is crucial so as not to systematically over-estimate travel distance, considering that a correlation exists between communicates rates and mobility. Efforts should be provided to address this bias, for instance by applying correction factors to the distance estimates. Additionally, we consider looking into ways to refine *a posteriori* the spatial and temporal resolution of these estimates by re-distributing the traffic volume in space and time.

- **Average traffic speed estimation** (Chapter 5): Estimating the temporal biases that communication behaviors imply in CDR users' observed travel times. We suggest resorting to the (temporally) regular activity chains so as to refine the identification of departure and arrival times with longitudinal analyses. This characterization step will support applying of traffic speed estimation method to real CDR data.

When it comes to the adapting the framework design for complete evaluation and on-field applications, the following research works still need to be conducted.

- First, in order to be fully operative, the framework requires interfacing with transport mode detection methods, that characterize individual modes and collective modal shares. It is a crucial step to estimate mode-specific traffic volumes and average traffic speeds, a key condition for estimating accurate traffic-related emissions.

- Second, more studies should be done on the longitudinal analysis in order to: 1. determine the need for periodic or continuous learning in the studied territory; and 2. adapt the framework to the detection of mobility anomalies. These studies will help design a software solution based on this framework that allows the automated integration and processing of new data.

The pursuit of these objectives should allow, in the long run, the complete exploitation of this framework on a long data horizon, first on the city of Santiago de Cali, Colombia, then on new cities.

# Bibliography


Lauren Alexander, Shan Jiang, Mikel Murga, and Marta C. González. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, 58:240 – 250, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2015.02.018. URL http://www.sciencedirect.com/science/article/pii/S0968090X1500073X. Big Data in Transportation and Traffic Engineering.

Essam Algizawy, Tetsuji Ogawa, and Ahmed El-Mahdy. Real-time large-scale map matching using mobile phone data. *ACM Trans. Knowl. Discov. Data*, 11(4), July 2017. ISSN 1556-4681. doi: 10.1145/3046945. URL https://doi.org/10.1145/3046945.

Maurice Armitage Cadavid, Elena Londoño Gomez, Uriel Dario Cancelado Sanchez, Guido Escobar Morales, and Diana Maria Perilla Galvis. Cali en cifras 2018-2019. Technical report, Departamento Administrativo de Planeacion, Alcadia de Santiago de Cali, 2019.

Fereshteh Asgari, Vincent Gauthier, and Monique Becker. A survey on human mobility and its applications, 2013.

Fereshteh Asgari, Alexis Sultan, Haoyi Xiong, Vincent Gauthier, and Mounim A. El-Yacoubi. Ct-mapper: Mapping sparse multimodal cellular trajectories using a multilayer transportation network. *Computing Research Repository - CORR*, abs/1604.06577, 2016. URL http://arxiv.org/abs/1604.06577.

Kay Axhausen and Tommy Gärling. Activity-based approaches to travel analysis: Conceptual frameworks, models and research problems. *Transport Reviews - TRANSP REV*, 12:323–341, 10 1992. doi: 10.1080/01441649208716826.

Danya Bachir. *Estimating Urban Mobility with Mobile Network Geolocation Data Mining*. PhD thesis, Télécom SudParis (Institut Mines-Télécom), Université Paris Saclay, 2019.

Danya Bachir, Vincent Gauthier, Mounim El Yacoubi, and Ghazaleh Khodabandelou. Using mobile phone data analysis for the estimation of daily urban dynamics. In *ITSC 2017 : 20th International Conference on Intelligent Transportation Systems*, pages 626 – 632, Yokohama, Japan, October 2017. IEEE Computer Society. doi: 10.1109/ITSC.2017.8317956. URL https://hal.archives-ouvertes.fr/hal-01745767.

Hillel Bar-Gera. Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from israel. *Transportation Research Part C: Emerging Technologies*, 15(6):380–391, Dec 2007. ISSN 0968-090X. doi: 10.1016/j.trc.2007.06.003. URL http://dx.doi.org/10.1016/j.trc.2007.06.003.

Albert-László Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–11, May 2005. doi: 10.1038/nature03459.

S. F. A. Batista, Ludovic Leclercq, and Monica Menendez. Dynamic traffic assignment for regional networks with traffic-dependent trip lengths and regional paths. *Transportation Research Part C: Emerging Technologies*, 2021a.

S.F.A. Batista, Ludovic Leclercq, and Nikolas Geroliminis. Estimation of regional trip length distributions for the calibration of the aggregated network traffic models. *Transportation Research Part B: Methodological*, 122:192 – 217, 2019. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2019.02.009. URL http://www.sciencedirect.com/science/article/pii/S0191261518311603.

S.F.A. Batista, Manon Seppecher, and Ludovic Leclercq. Identification and characterizing of the prevailing paths on a urban network for mfd-based applications. *Transportation Research Part C: Emerging Technologies*, 127:102953, 2021b. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102953. URL https://www.sciencedirect.com/science/article/pii/S0968090X20308500.

Tom Baum. Seasonality in tourism: Understanding the challenges: Introduction. *Tourism Economics*, 5(1):5–8, 2021/10/06 1999. doi: 10.1177/135481669900500101. URL https://doi.org/10.1177/135481669900500101.

Richard Becker, Ramón Cáceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, and C. Volinsky. A tale of one city: Using cellular network data for urban planning. *IEEE Pervasive Computing*, 10(4):18–26, April 2011. ISSN 1536-1268. doi: 10.1109/MPRV.2011.44.

Vincent Blondel, Adeline Decuyper, and Gautier Krings. A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4, 02 2015. doi: 10.1140/epjds/s13688-015-0046-0.

Loïc Bonnetain, Angelo Furno, Jean Krug, and Nour-Eddin El Faouzi. Can we map-match individual cellular network signaling trajectories in urban environments? data-driven study. *Transportation Research Record: Journal of the Transportation Research Board*, 2673:036119811984747, 05 2019. doi: 10.1177/0361198119847472.

Loïc Bonnetain, Angelo Furno, Nour-Eddin El Faouzi, Marco Fiore, Razvan Stanica, Zbigniew Smoreda, and Cezary Ziemlicki. Transit: Fine-grained human mobility trajectory inference at scale with mobile network signaling data. *Transportation Research Part C: Emerging Technologies*, 130:103257, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103257. URL https://www.sciencedirect.com/science/article/pii/S0968090X21002692.

Francesco Calabrese, Giusy Di Lorenzo, Liang Liu, and Carlo Ratti. Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in boston metropolitan area. *IEEE Pervasive Computing*, 10(4):36–44, 2011.

Julián Candia, Marta C. González, Pu Wang, Timothy Schoenharl, Greg Madey, and Albert-László Barabási. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 41(22):224015, June 2008. doi: 10.1088/1751-8113/41/22/224015. URL http://stacks.iop.org/1751-8121/41/i=22/a=224015?key=crossref.97d23b44de724a7398482cd45c7fe01a.

Pablo Samuel Castro, Daqing Zhang, Chao Chen, Shijian Li, and Gang Pan. From taxi gps traces to social and community dynamics: A survey. *ACM Comput. Surv.*, 46 (2), December 2013. ISSN 0360-0300. doi: 10.1145/2543581.2543584. URL https://doi.org/10.1145/2543581.2543584.

Cynthia Chen, Jingtao Ma, Yusak Susilo, Yu Liu, and Menglin Wang. The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, 2016.

Guangshuo Chen, Sahar Hoteit, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Enriching sparse mobility information in call detail records. *Computer Communications*, 2018.

Guangshuo Chen, Aline Carneiro Viana, Marco Fiore, and Carlos Sarraute. Complete trajectory reconstruction from sparse mobile phone data. *EPJ Data Science*, 8(1):30, 2019. doi: 10.1140/epjds/s13688-019-0206-8. URL https://doi.org/10.1140/epjds/s13688-019-0206-8.

Zhaoming Chu, Lin Cheng, and Hui Chen. A review of activity-based travel demand modeling. pages 48–59, 07 2012. ISBN 978-0-7844-1244-2. doi: 10.1061/9780784412442.006.

Cisco. Cisco annual internet report (2018-2023). Technical report, Cisco, 2020.

Serdar Çolak, Lauren P Alexander, Bernardo G Alvim, Shomik R Mehndiratta, and Marta C. González. Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. *Transportation research record: Journal of the transportation research board*, 2526(1):126–135, 2015.

Thomas Couronne, Zbigniew Smoreda, and Ana-Maria Olteanu. Chatty mobiles:individual mobility and communication patterns, 2013.

Carlos F. Daganzo. Urban gridlock: Macroscopic modeling and mitigation approaches. *Transportation Research Part B: Methodological*, 41(1):49 – 62, 2007. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2006.03.001. URL http://www.sciencedirect.com/science/article/pii/S0191261506000282.

DANE. Proyecciones de población.

Departamento Administrativo de Planeación Municipal. Proyecciones de poblacion de cali por comuna y corregimiento 2006-2020, 11 2018. URL http://datos.cali.gov.co/dataset/proyecciones-de-poblacion-de-cali-por-comuna-y-corregimiento-2006-2020.

Thierry Derrmann, Raphaël Frank, Francesco Viti, and T. Engel. Estimating urban road traffic states using mobile network signaling data. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7, Oct 2017. doi: 10.1109/ITSC.2017.8317718.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231, 1996.

Masoud Fallah Shorshani, Michel Andre, Céline Bonhomme, and Christian Seigneur. Modelling chain for the effect of road traffic on air and water quality: Techniques, current status and future prospects. *Environmental Modelling and Software*, 64:102–123, 02 2015. doi: 10.1016/j.envsoft.2014.11.020.

Mohammad Forghani, Farid Karimipour, and Christophe Claramunt. From cellular positioning data to trajectories: Steps towards a more accurate mobility exploration. *Transportation Research Part C: Emerging Technologies*, 117:102666, 2020. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102666. URL https://www.sciencedirect.com/science/article/pii/S0968090X20305817.

B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Analysis of gsm calls data for understanding user mobility behavior. In *2013 IEEE International Conference on Big Data*, pages 550–555, Oct 2013. doi: 10.1109/BigData.2013.6691621.

Barbara Furletti, Lorenzo Gabrielli, Chiara Renso, and Salvatore Rinzivillo. Identifying users profiles from mobile calls habits. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 17–24, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1542-5. doi: 10.1145/2346496.2346500. URL http://doi.acm.org/10.1145/2346496.2346500.

A. Furno, M. Fiore, R. Stanica, Cezary Ziemlicki, and Zbigniew Smoreda. A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE Transactions on Mobile Computing*, 16(10):2682–2696, Oct 2017. ISSN 1536-1233. doi: 10.1109/TMC.2016.2637901.

L. Gabrielli, B. Furletti, R. Trasarti, F. Giannotti, and D. Pedreschi. City users' classification with mobile phone data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1007–1012, Oct 2015. doi: 10.1109/BigData.2015.7363852.

Yerali Gandica, João Carvalho, Fernando Sampaio dos Aidos, Renaud Lambiotte, and Timoteo Carletti. Stationarity of the inter-event power-law distributions. *PLOS ONE*, 12(3):1–10, 03 2017. doi: 10.1371/journal.pone.0174509. URL https://doi.org/10.

1371/journal.pone.0174509.

Nikolas Geroliminis and Carlos F. Daganzo. Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759 – 770, 2008. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2008.02.002. URL http://www.sciencedirect.com/science/article/pii/S0191261508000180.

Marta C. Gonzalez, César A. Hidalgo, and Albert-László Barabási. Understanding individual human mobility patterns. *Nature*, 453:779 EP –, June 2008. URL https://doi.org/10.1038/nature06958.

Ramaswamy Hariharan and Kentaro Toyama. Project lachesis: Parsing and modeling location histories. In Max J. Egenhofer, Christian Freksa, and Harvey J. Miller, editors, *Geographic Information Science*, pages 106–124, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30231-5.

Sahar Hoteit, Guangshuo Chen, Aline C Viana, and Marco C Fiore. Spatio-Temporal Completion of Call Detail Records for Human Mobility Analysis. In *Rencontres Francophones sur la Conception de Protocoles, l'Évaluation de Performance et l'Expérimentation des Réseaux de Communication*, Quiberon, France, May 2017. URL https://hal.archives-ouvertes.fr/hal-01516717.

Haosheng Huang, Yi Cheng, and Robert Weibel. Transport mode detection based on mobile phone network data: A systematic review. *Transportation Research Part C: Emerging Technologies*, 101:297 – 312, 2019. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2019.02.008. URL http://www.sciencedirect.com/science/article/pii/S0968090X1831369X.

Infometrika. Comportamiento de las actividades de viajes y turismo en la ciudad de santiago de cali. Technical report, Alcadia de Santiago de Cali, Sep 2019.

Md. Shahadat Iqbal, Charisma F. Choudhury, Pu Wang, and Marta C. González. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies*, 40:63 – 74, 2014. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2014.01.002. URL http://www.sciencedirect.com/science/article/pii/S0968090X14000059.

Andreas Janecek, Danilo Valerio, Karin Anna Hummel, Fabio Ricciato, and Helmut Hlavacs. The cellular network as a sensor: From mobile phone data to real-time road traffic monitoring. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, 2015.

Shan Jiang, Gaston A. Fiore, Yingxiang Yang, Joseph Ferreira, Emilio Frazzoli, and Marta C. González. A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In *UrbComp@KDD*, 2013.

Shan Jiang, Joseph Ferreira, and Marta C. González. Activity-based human mobility patterns inferred from mobile phone data: A case study of singapore. *IEEE Transactions on Big Data*, 3(2):208–219, June 2017. ISSN 2332-7790. doi: 10.1109/TBDATA.2016.2631141.

Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. Understanding human mobility from twitter. *PLOS ONE*, 10(7):1–16, 07 2015. doi: 10.1371/journal.pone.0131469. URL https://doi.org/10.1371/journal.pone.0131469.

Zihan Kan, Luliang Tang, Mei-Po Kwan, Chang Ren, Dong Liu, Tao Pei, Yu Liu, Min Deng, and Qingquan Li. Fine-grained analysis on fuel-consumption and emission from vehicles trace. *Journal of Cleaner Production*, 203:340 – 352, 2018. ISSN 0959-6526. doi: https://doi.org/10.1016/j.jclepro.2018.08.222. URL http://www.sciencedirect.com/science/article/pii/S095965261832571X.

Ludovic Leclercq, Nicolas Chiabaut, and Béatrice Trinquier. Macroscopic fundamental

diagrams: A cross-comparison of estimation methods. *Transportation Research Part B: Methodological*, 2014.

Qing Li, Yang Cheng, Fan Ding, Xia Wan, and Bin Ran. Citywide hourly traffic emissions estimation using cellular activity data. In *TRB 95th Annual Meeting Compendium of Papers*, 2016.

Miao Lin and Wen-Jing Hsu. Mining gps data for mobility patterns: A survey. *Pervasive and Mobile Computing*, 12:1–16, 2014. ISSN 1574-1192. doi: https://doi.org/10.1016/j.pmcj.2013.06.005. URL https://www.sciencedirect.com/science/article/pii/S1574119213000825.

Jielun Liu, Ke Han, Ghim Ping Ong, et al. Spatial-temporal inference of urban traffic emissions based on taxi trajectories and multi-source urban data. *arXiv preprint arXiv:1809.10834*, 2018.

Ko Ko Lwin, Yoshihide Sekimoto, and Wataru Takeuchi. Estimation of hourly link population and flow directions from mobile cdr. *ISPRS International Journal of Geo-Information*, 7(11), 2018. ISSN 2220-9964. doi: 10.3390/ijgi7110449. URL https://www.mdpi.com/2220-9964/7/11/449.

Marco Mamei and Massimo Colonna. Analysis of tourist classification from cellular network data. *Journal of Location Based Services*, 12(1):19–39, 2018. doi: 10.1080/17489725.2018.1463466. URL https://doi.org/10.1080/17489725.2018.1463466.

Guilhem Mariotte, Ludovic Leclercq, S.F.A. Batista, Jean Krug, and Mahendra Paipuri. Calibration and validation of multi-reservoir mfd models: A case study in lyon. *Transportation Research Part B: Methodological*, 136:62 – 86, 2020. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2020.03.006. URL http://www.sciencedirect.com/science/article/pii/S0191261519306769.

Metro Cali. Encuesta de movilidad, 2015, 2015. URL https://www.metrocali.gov.co/wp/wp-content/uploads/2019/02/Encuesta-de-movilidad-2015.pdf.

Diala Naboulsi, Marco Fiore, Stephane Ribot, and Razvan Stanica. Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys Tutorials*, 18(1):124–161, 2016.

Andrew S. Nagle and Vikash V. Gayah. Accuracy of networkwide traffic states estimated from mobile probe data. *Transportation Research Record*, 2421(1):1–11, 2014. doi: 10.3141/2421-01. URL https://doi.org/10.3141/2421-01.

Mirco Nanni, Roberto Trasarti, Barbara Furletti, Lorenzo Gabrielli, Peter Van Der Mede, Joost De Bruijn, Erik De Romph, and Gerard Bruil. Transportation planning based on gsm traces: A case study on ivory coast. In Jordi Nin and Daniel Villatoro, editors, *Citizen in Sensor Networks*, pages 15–25, Cham, 2014. Springer International Publishing. ISBN 978-3-319-04178-0.

Kati Nilbe, Rein Ahas, and Siiri Silm. Evaluating the travel distances of events visitors and regular visitors using mobile positioning data: The case of estonia. *Journal of Urban Technology*, 21(2):91–107, apr 2014. doi: 10.1080/10630732.2014.888218. URL https://doi.org/10.1080%2F10630732.2014.888218.

Leonidas Ntziachristos, Dimitrios Gkatzoflias, Chariton Kouridis, and Zissis Samaras. Copert: A european road transport emission inventory model. In Ioannis N. Athanasiadis, Andrea E. Rizzoli, Pericles A. Mitkas, and Jorge Marx Gómez, editors, *Information Technologies in Environmental Engineering*, pages 491–504, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg. ISBN 978-3-540-88351-7.

Marguerite Nyhan, Stanislav Sobolevsky, Chaogui Kang, Prudence Robinson, Andrea Corti, Michael Szell, David Streets, Zifeng Lu, Rex Britter, Steven Barrett, and Carlo Ratti. Predicting vehicular emissions in high spatial resolution using pervasively measured transportation data and microscopic emissions model. *Atmospheric Environment*, 140, 06 2016. doi: 10.1016/j.atmosenv.2016.06.018.

Joaquín Osorio-Arjona and Juan Carlos García-Palomares. Social media and urban mo-

bility: Using twitter to calculate home-work travel matrices. *Cities*, 89:268 – 280, 2019. ISSN 0264-2751. doi: https://doi.org/10.1016/j.cities.2019.03.006. URL http://www.sciencedirect.com/science/article/pii/S0264275118312976.

Q. Ou, R. L. Bertini, J. W. C. van Lint, and S. P. Hoogendoorn. A theoretical framework for traffic speed estimation by fusing low-resolution probe vehicle data. *IEEE Transactions on Intelligent Transportation Systems*, 12(3):747–756, Sep. 2011. ISSN 1558-0016. doi: 10.1109/TITS.2011.2157688.

M. Paipuri, E. Barmpounakis, N. Geroliminis, and L. Leclercq. Linear regression analysis of regional mean speed of athens city network using drone data: A multi-modal approach. In *100th TRB Annual Meeting*, 2021.

Mahendra Paipuri, Yanyan Xu, Marta C. González, and Ludovic Leclercq. Estimating mfds, trip lengths and path flow distributions in a multi-region setting using mobile phone data. *Transportation Research Part C: Emerging Technologies*, 118:102709, 2020. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2020.102709. URL http://www.sciencedirect.com/science/article/pii/S0968090X20306240.

Mozhgan Pourmoradnasseri, Kaveh Khoshkhah, Artjom Lind, and Amnir Hadachi. Od-matrix extraction based on trajectory reconstruction from mobile data. In *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, pages 1–8, 2019. doi: 10.1109/WiMOB.2019.8923358.

Gyan Ranjana, Hui Zang, Zhi-Li Zhang, and Jean Bolot. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16:33–44, 12 2012. doi: 10.1145/2412096.2412101.

Daniel Rodriguez-Rey, Marc Guevara, M Paz Linares, Josep Casanovas, Juan Salmerón, Albert Soret, Oriol Jorba, Carles Tena, and Carlos Pérez García-Pando. A coupled macroscopic traffic and pollutant emission modelling system for barcelona. *Transportation Research Part D: Transport and Environment*, 92:102725, 2021. ISSN 1361-9209. doi: https://doi.org/10.1016/j.trd.2021.102725. URL https://www.sciencedirect.com/science/article/pii/S1361920921000274.

Mario B. Rojas, Eazaz Sadeghvaziri, and Xia Jin. Comprehensive review of travel behavior and mobility pattern studies that used mobile phone data. *Transportation Research Record Journal of the Transportation Research Board*, January 2016.

Christos Samaras, Dimitris Tsokolis, Silvana Toffolo, Giorgio Magra, Leonidas Ntziachristos, and Zissis Samaras. Improving fuel consumption and co 2 emissions calculations in urban areas by coupling a dynamic micro traffic model with an instantaneous emissions model. *Transportation Research Part D: Transport and Environment*, 65, 11 2017. doi: 10.1016/j.trd.2017.10.016.

Secretaría de Turismo de Cali. Boletín de estadísticas de turismo. Technical report, Alcadia de Santiago de Cali, 2019.

Manon Seppecher, Ludovic Leclercq, Angelo Furno, Delphine Lejri, and Thamara Vieira da Rocha. Estimation of urban zonal speed dynamics from user-activity-dependent positioning data and regional paths. *Transportation Research Part C: Emerging Technologies*, 129:103183, 2021. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2021.103183. URL https://www.sciencedirect.com/science/article/pii/S0968090X21001996.

Jingbo Shang, Yu Zheng, Wenzhu Tong, Eric Chang, and Yong Yu. Inferring gas consumption and pollution emissions of vehicles throughout a city. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug 2014.

Ratul Sikder, Md. Jamal Uddin, and Sajal Halder. An efficient approach of identifying tourist by call detail record analysis. In *2016 International Workshop on Computational Intelligence (IWCI)*, pages 136–141, 2016. doi: 10.1109/IWCI.2016.7860354.

Robin Smit, Muriel Poelman, and Jeroen Schrijver. Improved road traffic emission inventories by adding mean speed distributions. *Atmospheric Environment*, 42(5):916–926,

2008. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2007.10.026. URL https://www.sciencedirect.com/science/article/pii/S1352231007009041.

Robin Smit, Leonidas Ntziachristos, and Paul Boulter. Validation of road vehicle and traffic emission models – a review and meta-analysis. *Atmospheric Environment*, 44(25): 2943–2953, 2010. ISSN 1352-2310. doi: https://doi.org/10.1016/j.atmosenv.2010.05.022. URL https://www.sciencedirect.com/science/article/pii/S135223101000395X.

Chaoming Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, Feb 2010. ISSN 1095-9203. doi: 10.1126/science.1177170. URL http://dx.doi.org/10.1126/science.1177170.

Etienne Thuillier, Laurent Moalic, Sid Ahmed Lamrous, and Alexandre Caminada. Clustering Weekly Patterns of Human Mobility Through Mobile Phone Data. *IEEE Transactions on Mobile Computing*, 17(4):817–830, 2018. doi: 10.1109/TMC.2017.2742953. URL https://hal.archives-ouvertes.fr/hal-01992673.

Eran Toch, Boaz Lerner, Eyal Ben-Zion, and Irad Ben-Gal. Analyzing large-scale human mobility data: a survey of machine learning methods and applications. *Knowledge and Information Systems*, Mar 2018. ISSN 0219-3116. doi: 10.1007/s10115-018-1186-x. URL https://doi.org/10.1007/s10115-018-1186-x.

Jameson L. Toole, Serdar Çolak, Bradley Sturt, Lauren P. Alexander, Alexandre Evsukoff, and Marta C. González. The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C: Emerging Technologies*, 58:162 – 177, 2015. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2015.04.022. URL http://www.sciencedirect.com/science/article/pii/S0968090X15001631. Big Data in Transportation and Traffic Engineering.

Nikolaos Tsanakas, Joakim Ekström, and Johan Olstam. Estimating emissions from static traffic models: Problems and solutions. *Journal of Advanced Transportation*, 2020:1–17, 2020.

Maarten Vanhoof, Liane Hendrickx, Aare Puussaar, Gert Verstraeten, Thomas Ploetz, and Zbigniew Smoreda. Exploring the use of mobile phone data for domestic tourism trip analysis. *Netcom*, 31:335–372, 12 2017. doi: 10.4000/netcom.2742.

Maarten Vanhoof, Fernando Reis, Zbigniew Smoreda, and Thomas Ploetz. Detecting home locations from cdr data: introducing spatial uncertainty to the state-of-the-art, 08 2018.

Thamara Vieira da Rocha, Arnaud Can, Céline Parzani, Bruno Jeanneret, Rochdi Trigui, and Ludovic Leclercq. Are vehicle trajectories simulated by dynamic traffic models relevant for estimating fuel consumption? *Transportation Research Part D: Transport and Environment*, 24:17–26, 2013. ISSN 1361-9209. doi: https://doi.org/10.1016/j.trd.2013.03.012. URL https://www.sciencedirect.com/science/article/pii/S1361920913000618.

Hai Yang, Jintao Ke, and Jieping Ye. A universal distribution law of network detour ratios. *Transportation Research Part C: Emerging Technologies*, 96:22–37, 2018. ISSN 0968-090X. doi: https://doi.org/10.1016/j.trc.2018.09.012. URL https://www.sciencedirect.com/science/article/pii/S0968090X18311185.

Mehmet Yildirimoglu and Nikolas Geroliminis. Approximating dynamic equilibrium conditions with macroscopic fundamental diagrams. *Transportation Research Part B: Methodological*, 70:186 – 200, 2014. ISSN 0191-2615. doi: https://doi.org/10.1016/j.trb.2014.09.002. URL http://www.sciencedirect.com/science/article/pii/S0191261514001568.

X. Zhan, Y. Zheng, X. Yi, and Satish V. Ukkusuri. Citywide traffic volume estimation using trajectory data. *IEEE Transactions on Knowledge and Data Engineering*, 29(2): 272–285, Feb 2017. ISSN 1041-4347. doi: 10.1109/TKDE.2016.2621104.

Junping Zhang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen. Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Trans-*

*portation Systems*, 12:1624–1639, 12 2011. doi: 10.1109/TITS.2011.2158001.

Zhan Zhao, Haris N. Koutsopoulos, and Jinhua Zhao. Identifying hidden visits from sparse call detail record data, 2021.

Vincent Zheng, Yu Zheng, Xing Xie, and Qiang Yang. *Collaborative location and activity recommendations with GPS history data*. 01 2010. doi: 10.1145/1772690.1772795.

Yu Zheng, Furui Liu, and Hsun-Ping Hsieh. U-air: When urban air quality inference meets big data. August 2013. URL https://www.microsoft.com/en-us/research/publication/u-air-when-urban-air-quality-inference-meets-big-data/.

Yu Zheng, Yi Xiuwen, Ming Li, Ruiyuan Li, Zhangqing Shan, Eric Chang, and Tianrui Li. Forecasting fine-grained air quality based on big data. In *KDD'15*, 2015.

# Appendices

# Appendix A

# Appendix for Chapter 2

## A.1 Resident scaling factors

| Admin. division | Scaling factor |
|---|---|
| Comuna 1 | 6.735974 |
| Comuna 2 | 2.882930 |
| Comuna 3 | 3.445209 |
| Comuna 4 | 3.426846 |
| Comuna 5 | 3.711358 |
| Comuna 6 | 4.926330 |
| Comuna 7 | 3.669392 |
| Comuna 8 | 3.568015 |
| Comuna 9 | 3.228763 |
| Comuna 10 | 3.633109 |
| Comuna 11 | 4.538028 |
| Comuna 12 | 5.614626 |
| Comuna 13 | 4.714124 |
| Comuna 14 | 4.607423 |
| Comuna 15 | 5.136440 |
| Comuna 16 | 5.455057 |
| Comuna 17 | 2.807282 |
| Comuna 18 | 4.316421 |
| Comuna 19 | 3.137365 |
| Comuna 20 | 6.391412 |
| Comuna 21 | 4.267734 |
| Comuna 22 | 1.090481 |
| Los Andes | 3.270313 |
| El Saladito | 1.425073 |
| Golondrinas | 2.582787 |
| Montebello | 2.301922 |
| La Castilla | 1.335343 |
| Villacarmelo | 9.494643 |
| La Buitrera | 1.833327 |
| Felidia | 4.219613 |
| La Elvira | 0.985164 |
| La Paz | 8.281481 |
| La Leonera | 4.566387 |
| Pichinde | 4.780380 |
| El Hormiguero | 0.533089 |
| Navarro | 3.860305 |
| Pance | 1.383871 |
| Yumbo | 2.889096 |
| Jamundi | 2.569953 |
| Commuters | 1.426534 |
| Visitors | 0.270245 |

Table A.1: Table of scaling factors by administrative division

# Appendix B

# Appendix for Part II

## B.1 Introduction

Extracting the mobility patterns from the communication data is a crucial step that supports all the mobility analyses. This Chapter presents the extraction process implemented in this thesis.

The CDR-based literature has adapted existing methods initially developed for GPS data processing [Hariharan and Toyama, 2004]. These adaptations result in a general algorithm proven to identify static phases of mobility from communication events [Jiang et al., 2013, Toole et al., 2015]. Although there are variations in their implementation, these methods generally follow the same data processing chain.

1. Identifying *stays*, *i.e.*, events generated within a bounded fixed distance with a minimum duration. The identification of these stays can be based on the roaming distance as implemented here, or on more elaborated approach like hierarchical clustering [Jiang et al., 2013, Toole et al., 2015].

2. Stabilizing of close stays, considering that they correspond actually to the same location visited. Again, this calls on aggregate methods. We resort to a greedy grid-based approach [Jiang et al., 2013, Zheng et al., 2010]. A more simple alternative is provided in [Toole et al., 2015].

3. Identifying potential stays, *i.e.*, the remaining communication events that colocate with stay positions (without being identified as such for lack of sufficient time) are labeled as potential stays.

4. Labeling of remaining communication sequences as pass-by points.

This approach has proven useful for extracting individual or collective mobility patterns at the urban scale, especially in works aiming at reconstructing Origin-Destination matrices. However However, this literature has so far mainly focused on resident populations, thus evacuating the issue of detecting flows into and out of the area. Our ambition is to consider dynamic populations characterized by inner and outer flows, like commuters and visitors. In that case, the question arises of knowing whether this stay detection method can capture the movement of such populations. Considering that commuting users perform a significant share of their stays out of the area, it is likely that the trips from and to those external locations will be missed by simply using the literature approach.

Therefore, we propose an extension of the literature method that targets the mobility of such users. Based on the observation of inactivity rather than the activity phase, we define quasi-stays as a meaningful location of a CDR track without asserting users spent time there.

## B.2 Method

Let us consider a hypothetical user with a commuting mobility profile. Their activity within the city can be detected based on the stay and potential stay detection methods. On this basis, the inner trips, such as the one between the workplace and a leisure location, can be detected. However, the home location being outside of the monitored area, all the trips departing from or reaching out to this location will be undetected. It can result in a significant underestimation of distance traveled and related emissions, especially for peak hours.

Let us now consider outgoing work-home trip for a user working in the area and living outside of it. While their destination (home location) will remain unknown, some pass-by-points along the trip can play the role of destination proxy, which we will call *quasi-stays*. It is relevant to try to detect the last of these pass-by points: it is the one that will be the closest to the area boundary and therefore will best characterize the distances traveled. Among the pass-by points identified along the trip, we should therefore look for the one that is followed by a significant period of inactivity, during which the user is assumed to be outside of the area. Unlike the stay detection described above, which consists in identifying sufficiently long phases of communication activity, we propose this time to identify sufficiently long phases of inactivity (1 hour). In order to target more precisely this kind of movement, we additionally propose to set a minimum distance that should be respected between the last observed stay and the considered pass-by-point event. If those two criteria are satisfied, then the pass-by-point is re-labeled as a quasi-stay.

This algorithm is specially adapted to detect outgoing quasi-stays (long periods of inactivity far enough from the preceding stay). We apply the symmetrical method to detect incoming quasi-stays, *i.e.* pass-by-points preceded by long periods of inactivity far enough from the next stay. The following section comments on the results obtained.

## B.3 Impact assessment

First, we focus on the analysis of the results at an individual level. Figures B.1a and B.1d represent their mobility pattern based on every mobility events over a historical period of two months (January and February 2021). The user appears to mainly commute between Cali's city center and a border cell. Being associated with a commuting mobility profile, they very likely come from the neighboring municipality of Palmira. Figures B.1b and B.1e represent the mobility pattern as detected from M1. We observe that with this approach, an essential share of the commuting trips is not detected. Figures B.1c and B.1f represent the mobility pattern as detected from M2. It happens because not enough events occur at that position to define a stay. However, these events are followed by long inactivity phases and are sufficiently far from the next stay, allowing them to be considered quasi-stays.

Interestingly, the method also allows detecting trips towards the municipality of Yumbo, in the North of Cali, where the user was not sufficiently active for stays to be detected there. It indicates that the meaningful events we detect should be carefully qualified as they do not necessarily and clearly correspond to border crossings. In this case, the quasi-stay detected in Yumbo could be the last event on the way to the northern area border. However, it could also be an activity carried out in Yumbo without sufficient events to allow its labeling to stay. In any case, including this quasi-stay in the mobility analysis will necessarily return better flow and distances result than not.

This observation raises the issue of the sensitivity of the method to the last (or first) detected pass-by-point. It could occur at any stage of the trip and in different positions from a day to another. Selecting the most frequent locations or the further away from the previous stay could be a way to stabilize and make quasi-stays converge around a few

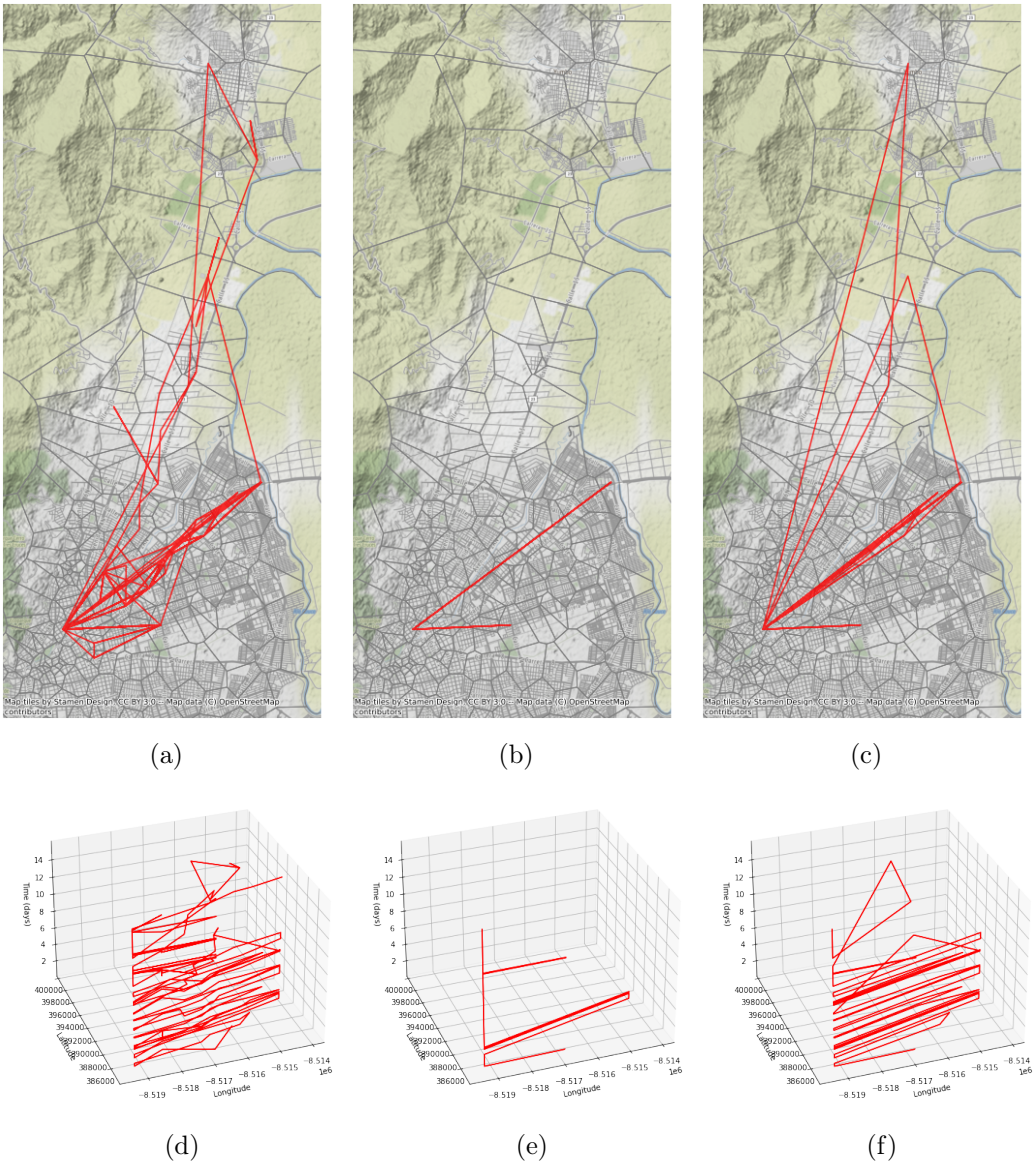
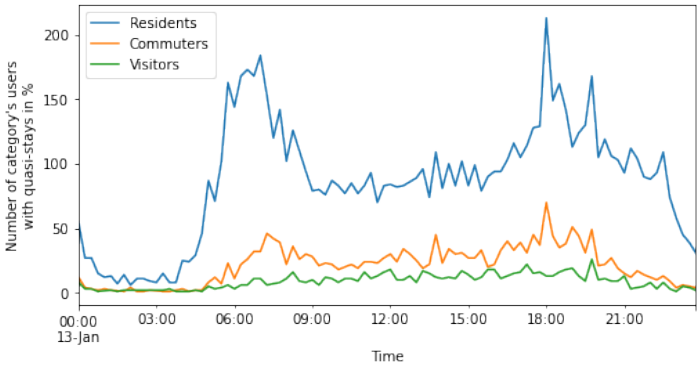



Figure B.1: Comparision of stay detection using M1 and M2, through the observation of spatial and spatio-temporal patterns.

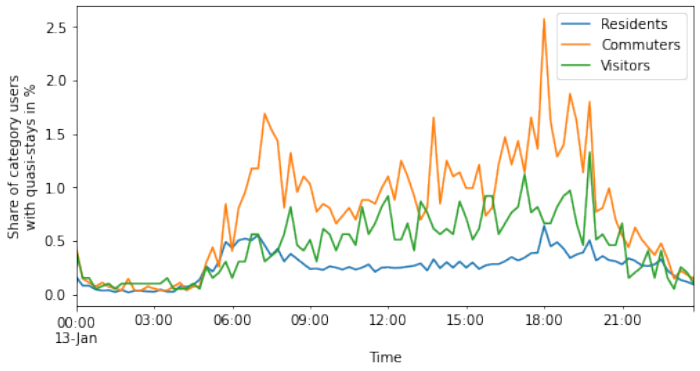
locations only. It may be the subject of later improvements of this method.

We measure which user category is the most impacted by the method change and how at the population scale. Figure B.2 displays the number of quasi-stays detected over a standard mobility day per user category. Figure B.2a illustrate the absolute number of quasi-stays detected over the day. Instead, Figure B.2b illustrates the number of quasi-stays divided by the observed category size. The residents generate a larger share of the quasi-stays detected. However, the proportion of users generating quasi-stays among each category is at least twice more significant for commuters than for residents. We observe that the temporal trend of quasi-stays detected over the day displays peak hours in the morning and late afternoon and a smaller peak at noon. The resident trend is also influenced by peak hours in the morning and afternoon. Our previous observation explains this: the method is helpful to detect entries and exits and enhance the detection of significant visited places where users are not so active. The visitors have a less characteristic profile, which is not surprising since the mobility patterns are likely to have more diffuse behaviors.

In Figure B.3, we also observe the spatial distribution of quasi-stays detected for the




(a) Number of QS detected




(b) Number of QS detected
normed with observed pop. size

Figure B.2: Impact assessment of the quasi stay detection

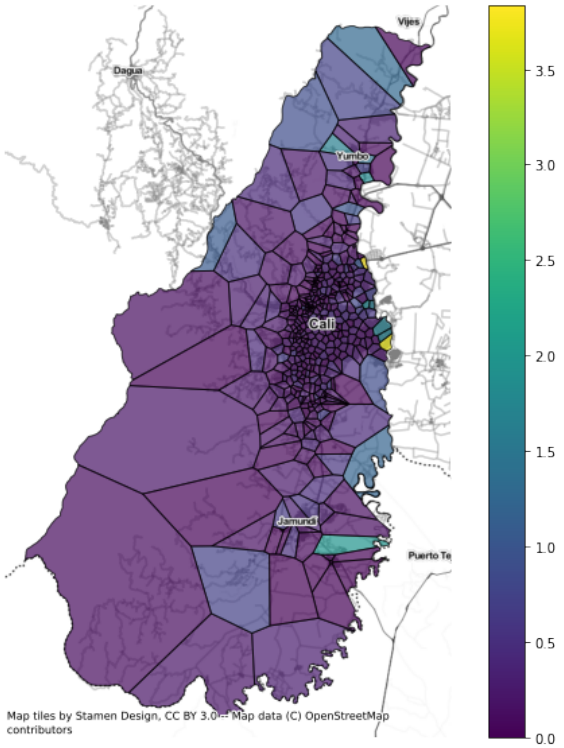



Figure B.3: Spatial repartition of quasi stays for the commuting population

commuting users. The color gradient indicates the percentage of quasi-stays generated in a considered BS cell. Interestingly, the cells that concentrate the largest quasi-stays are located either on the area border or in Cali's neighboring Yumbo and Jamundi. In particular, to cell concentrate each more than 3.5% of the quasi-stays on the area's eastern border. Both intersect principal road axes between Cali and the external neighbor municipalities of Candelaria and Palmira. It confirms the ability of our method to detect entrance and exit communication events. We can interpret the events massively detected in Yumbo and Jamundi in several ways. Either a share of those users are wrongly considered as external commuters and are, in reality, residents of Yumbo and Jamundi; It can also show that users tend to carry out in those area activities during which they do not use their phones (cf. user mobility pattern displayed in Figure B.1 for instance). Finally, there might also be a communication bias leading commuters to generate their first or last event when they reach these towns rather than elsewhere on their way. As mentioned previously, the categorization of users will be iteratively improved. It should help to remove these uncertainties.

## B.4 Conclusions and perspectives

The CDR-related literature has successfully adapted methods of the GPS-based literature to identify static phases within CDR mobility tracks. Such methods were constructive to build up OD matrices from such data at the urban scale and based on the analysis of assumed residents. However, the enlargement of the application of these methods to other categories of users, especially external commuters, showed limits as these users happen to spend a part of their stays out of the monitored area. Therefore, this chapter proposes a method for identifying meaningful locations from the mobility perspective that might not necessarily correspond to a stay location. Instead of relying on the activity observation, we rather rely on the inactivity observation to identify a moment where users might have

left the area. In order to consider only the most significant locations, we focus on events followed (respectively preceded) by prolonged inactivity periods, and sufficiently afar from the last (respectively next) known stay.

Applying the method to our case study shows satisfactory results. At an individual stay, it has proven capable to synthesize the mobility pattern better than the literature approach. At the population level, the results show that the method has the most significant impact on the commuting population, which is the one targeted. It also has the most decisive impact on early morning and late afternoon events, consistent with our expectations. Finally, the spatial analysis of the distribution of quasi-stays shows that a significant share of such events originated from the border BS cells, especially those located on the central road axis towards the neighboring municipalities.

Although the detected quasi-stays may not capture the exact stay location of commuting users, one can consider the quasi-stay as a spatial projection of the latent stay on the monitored area. For those users, this can be a good proxy for the missing origin and destinations and allow us to approximate the mobility patterns and better estimate flows and traveled length than without it.

# Appendix C

# Appendix for Chapter 5

## C.1 Table of notations

Table C.1 summarizes the notation used in this paper.

Table C.1: Nomenclature used in this paper.

| | |
|---|---|
| *General notations:* | |
| $P$ | Generic regional path |
| $r$ | Generic region |
| $t$ | Generic time period |
| *Individual trip characteristics:* | |
| $i$ | Generic individual trip |
| $t_0^i$ | Actual arrival time of trip $i$ |
| $t_{0,obs}^i$ | Observed arrival time of trip $i$ |
| $t^i$ | Actual arrival period of trip $i$ |
| $t_{obs}^i$ | Observed arrival period of trip $i$ |
| $T_{(P)}^i$ | Actual travel time of trip $i$ (along $P$) |
| $T_{(P),obs}^i$ | Observed travel time of trip $i$ (along $P$) |
| $\epsilon_d^i$ | Temporal bias of trip $i$ existing between observed departure time and actual one |
| $\epsilon_a^i$ | Temporal bias of trip $i$ existing between actual arrival time and observed one |
| $\epsilon^i$ | Travel time bias on trip $i$ |
| $V_r^i$ | Average speed of $i$ in region $r$ |
| $L_{P,r}^i$ | Distance traveled in region $r$ of $P$ by $i$ |
| *Travel time estimation:* | |
| $I_P^t$ | Overlapping trips along $P$ reaching destination at $t$ |
| $n_{t,P}$ | Number of trips in $I_P^t$ |
| $\bar{T}_P^t$ | Average actual travel time of trips in $I_P^t$ |
| $\bar{T}_{P,obs}^t$ | Average observed travel time of trips in $I_P^t$ |
| $\bar{T}_{P,obs}^t$ | Average travel time bias of trips in $I_P^t$ |
| $T_{P,r}^i$ | Actual travel time of trip $i$ in region $r$ |
| $\bar{\epsilon}_P^t$ | Average bias of trips in $I_P^t$ |

*Continued on next page*

Table C.1 – *Continued from previous page*

| | |
|---|---|
| *Speed estimation:* | |
| $V_r^t$ | Mean spatial speed in region $r$ |
| $\bar{L}_{P,r}^t$ | Average distance traveled in $r$ along $P$ during period $t$ |
| $\hat{L}_{P,r}$ | Regional trip length estimate in region $r$ along $P$ |
| $x_r^t$ | Reciprocal of $V_r^t$ |
| $S^t$ | Equation system at period $t$ |
| $\hat{\boldsymbol{L}}$ | Trip length matrix estimate |
| $\hat{\boldsymbol{L}}^{\|t}$ | Sub-matrix of $\hat{\boldsymbol{L}}$ made of the regional paths observed at period $P$ |
| $\boldsymbol{T}_{\boldsymbol{obs}}^{\boldsymbol{t}}$ | average observed travel time vector |
| $\boldsymbol{x}_{\boldsymbol{0}}^{\boldsymbol{t}}$ | Solution vector of $S^t$ doing a least square regression |
| *Bias modeling:* | |
| $Z$ | random variable modeling the inter-event time distribution |
| $Y$ | random variable modeling the arrival and departure biases distributions |
| $X$ | random variable modeling the travel time bias distribution |
| $\mu_X$ | Average travel time bias estimate |

## C.2 Bias characterization

We detail here the calculation leading to the results in Section 5.3.1. In that section, we defined :

$$Z \sim Exp(\lambda) \tag{C.1}$$
$$Y|Z \sim U(0, z) \tag{C.2}$$

Marginalizing over $Z$, the probability density function of $Y$ can be expressed as:

$$f_Y(y) = \int_0^{+\infty} f_{Y|Z}(y \mid z) \cdot f_Z(z) dz \tag{C.3}$$
$$= \int_y^{+\infty} \frac{1}{z} \cdot \lambda e^{-\lambda z} dz \tag{C.4}$$
$$= \lambda \int_0^{+\infty} \frac{e^{-\lambda(y+z)}}{y+z} dz \tag{C.5}$$

The expected value of $Y$ is then calculated as follows:

$$E(Y) = \int_0^{+\infty} E(Y|Z=z) \cdot f_Z(z)dz \tag{C.6}$$
$$= \int_0^{+\infty} \frac{z}{2} \cdot f_Z(z)dz \tag{C.7}$$
$$= \frac{1}{2}\int_0^{+\infty} z \cdot f_Z(z)dz \tag{C.8}$$
$$= \frac{1}{2}E(Z) \tag{C.9}$$
$$= \frac{1}{2\lambda} \tag{C.10}$$

While the variance of $Y$ is given by:

$$V(Y) = E(Y^2) - E(Y)^2 \tag{C.11}$$

Yet:

$$E(Y^2) = \int_0^{+\infty} E(Y^2|Z=z) \cdot f_Z(z)dz \tag{C.12}$$
$$= \int_0^{+\infty} \frac{z^2}{3} \cdot f_Z(z)dz \tag{C.13}$$
$$= \frac{1}{3}E(Z^2) = \frac{1}{3}(V(Z) + E(Z)^2) \tag{C.14}$$
$$= \frac{1}{3}(\frac{1}{\lambda^2} + \frac{1}{\lambda^2}) \tag{C.15}$$
$$= \frac{2}{3\lambda^2} \tag{C.16}$$

Thus:

$$V(Y) = E(Y^2) - E(Y)^2 \tag{C.17}$$
$$= \frac{2}{3\lambda^2} - \frac{1}{4\lambda^2} \tag{C.18}$$
$$= \frac{5}{12}\frac{1}{\lambda^2} \tag{C.19}$$

## C.3 Trip Length Matrix Variation with time

In Figure C.1, we display the boxplot describing for each region the distribution of the relative errors between regional trip lengths computed on February's data and March's data. We observe larger errors in urban regions (Regions 0 to 9), while the ring road regions display lower ones. This observation is related to the fact that the trip lengths distances on the ring road are very constrained by the ring road linear structure. On the contrary, a given regional path has a more extensive range of regional trip lengths in the city center, explaining the larger errors. However, the errors are still bounded in the urban regions, which confirms a regularity of regional average trip lengths overtime. This observation supports our framework, as it guarantees that average trip lengths estimated from another period of time, possibly from an independent dataset, will still provide a reliable database for the speed estimation process.

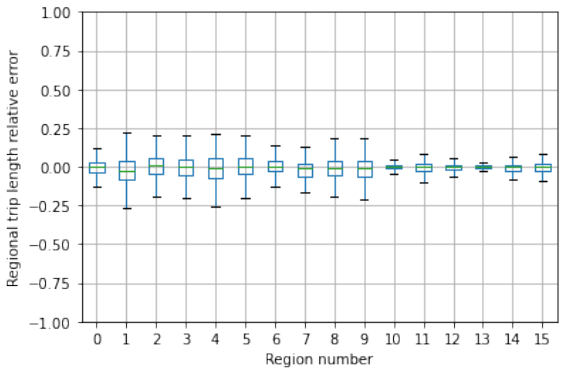



Figure C.1: Box plot of the relative errors of the regional trip lengths by region.